DOCUMENT RESUME

ED 414 340 TM 027 880

AUTHOR	Colton, Dean A.; Gao, Xiaohong; Harris, Deborah J.; Kolen, Michael J.; Martinovich-Barhite, Dara; Wang, Tianyou; Welch, Catherine J.
TITLE	Reliability Issues with Performance Assessments: A Collection of Papers. ACT Research Report Series 97-3.
INSTITUTION	American Coll. Testing Program, Iowa City, IA.
PUB DATE	1997-08-00
NOTE	137p.; The paper by Welch and Martinovich-Barhite was presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997), and versions of the other five papers were presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE	Collected Works - General (020) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	*Decision Making; Error of Measurement; Item Response Theory; *Performance Based Assessment; *Test Reliability
IDENTIFIERS	Bootstrap Methods; Polytomous Items; Weighting (Statistical)

ABSTRACT

This collection consists of six papers, each dealing with some aspects of reliability and performance testing. Each paper has an abstract, and each contains its own references. Papers include: (1) "Using Reliabilities To Make Decisions" (Deborah J. Harris); (2) "Conditional Standard Errors, Reliability, and Decision Consistency Performance Levels Using Polytomous IRT" (item response theory) (Tianyou Wang, Michael J. Kolen, and Deborah J. Harris); (3) "Assessing the Reliability of Performance Level Scores Using Bootstrapping" (Dean A. Colton, Xiaohong Gao, and Michael J. Kolen); (4) "Evaluating Measurement Precision of Performance Assessment with Multiple Forms, Raters, and Tasks" (Xiaohong Gao and Dean A. Colton); (5) "Weights that Maximize Reliability under a Congeneric Model for Performance Assessment" (Tianyou Wang); and (6) "Reliability Issues and Possible Solutions" (Catherine J. Welch and Dara Martinovich-Barhite). (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ACT Research Report Series 97-3

ED 414 340

Reliability Issues With Performance Assessments: A Collection of Papers

Dean A. Colton
Xiaohong Gao
Deborah J. Harris
Michael J. Kolen
Dara Martinovich-Barhite
Tianyou Wang
Catherine J. Welch

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM027880

ACT

BEST COPY AVAILABLE

August 1997

For additional copies write:
ACT Research Report Series
PO Box 168
Iowa City, Iowa 52243-0168

© 1997 by ACT, Inc. All rights reserved.

Reliability Issues With Performance Assessments: A Collection of Papers

Dean A. Colton
Xiaohong Gao
Deborah J. Harris
Michael J. Kolen
Dara Martinovich-Barhite
Tianyou Wang
Catherine J. Welch

Table of Contents

Page

Introduction ... iii

Using Reliabilities to Make Decisions ... 1
Deborah J. Harris

Conditional Standard Errors, Reliability, and Decision Consistency Performance Levels Using Polytomous IRT ... 13
Tianyou Wang, Michael J. Kolen, Deborah J. Harris

Assessing the Reliability of Performance Level Scores Using Bootstrapping ... 41
Dean A. Colton, Xiaohong Gao, Michael J. Kolen

Evaluating Measurement Precision of Performance Assessment With Multiple Forms, Raters, and Tasks ... 57
Xiaohong Gao, Dean A. Colton

Weights That Maximize Reliability Under a Congeneric Model for Performance Assessment ... 77
Tianyou Wang

Reliability Issues and Possible Solutions ... 95
Catherine J. Welch, Dara Martinovich-Barhite

Introduction

This report consists of six papers, each dealing with some aspect of reliability and performance testing. One of the papers, Welch and Martinovich-Barhite, was presented at the 1997 Annual Meeting of the American Educational Research Association in a symposium called *Issues in Large-Scale Portfolio Assessment*. Versions of the other five papers were presented at the 1996 Annual Meeting of the American Educational Research Association as part of a symposium called *Technical Issues Involving Reliability and Performance Assessments*. The authors would like to thank the discussants of the two symposia, Ed Wolfe, and Robert L. Brennan and Nancy L. Allen, respectively, for their comments during the two sessions, and Bradley A. Hanson and E. Matthew Schulz for their comments on a draft report.

Using Reliabilities to Make Decisions

Deborah J. Harris

Abstract

For a variety of reasons, there has been an increased use of performance assessments in high stakes and/or large scale situations. There is a long history of using performance assessments for classroom measurement; however, using these types of assessments beyond a single classroom (where a single administration has more long term consequences than whether to reteach the previous day's lesson) leads to an increased need for valid, reliable assessments. Validity and reliability issues relating to performance assessments have been much discussed, but further research and technical development is needed. For example, reliability with performance assessments has frequently been relegated to solely the agreement among the raters scoring the assessments. Although this is certainly an important component, it is not sufficient to ensure a reliable assessment.

This paper addresses the use of reliability information, such as that provided in the later papers in this report, in decision making. Specifically, choosing a score scale, forming a composite score, choosing a cut score, selecting a test, and similar issues are briefly discussed. Making a decision by choosing the highest reliability estimate does not always appear to be the optimal decision, particularly when reliability can be assessed in different ways (e.g., rater agreement, generalizability coefficient, using a theoretical model, or bootstrapping), and when the typical reliability estimate used with performance assessments, rater agreement, may not be the most relevant, given the purpose of the assessment.

The author would like to thank Michael J. Kolen and Catherine J. Welch for their comments on an earlier draft.

Using reliabilities to make decisions

There appears to be nearly universal agreement that reliability is an important property in measurement. Although the validity versus reliability argument may rage on in some quarters, few professionals seem to be arguing that reliability in and of itself is not a desirable property for a measurement instrument.

Nearly all technical manuals seem to report some sort of reliability value, and generally more than one. When a new method of testing is proposed, reliability is one of the first properties users want information about. The difficulty with reliability, therefore, lies not in the fact that it is not viewed as a valuable property, but in that there is no clear consensus as to the definition of reliability, or what it means, or what to do with reliability estimates. This difficulty appears more of a problem with performance assessments than it has in the past with multiple choice tests for various reasons.

Multiple choice tests can be made very reliable. Lengthening multiple choice tests to increase reliability is generally practical. Increasing reliability by lengthening the test also tends to increase some types of validity in that more items tend to more adequately cover the domain of interest. The various ways of defining/measuring reliability are less at odds in multiple choice testing. It is possible to develop a well-defined table of specifications, and to construct reasonably interchangeable forms from it, which not only are comparable to each other, but which also serve to cover the domain of interest reasonably well.

With performance assessments, increasing reliability may mean limiting the domain coverage either by constraining the domain itself or through more highly structuring responses, which may be at odds with how validity is viewed. Increasing the length of the test is more problematic than with multiple choice tests. Although it is certainly possible to develop a well defined table of specifications for performance assessments, there may be too little time available for testing to adequately cover the table of specifications in each form.

Test/retest or parallel forms reliability estimates are easier to obtain with multiple choice tests than with most performance assessments, because of the time involved and because of the possible lack of truly comparable performance assessment forms. In some instances, such as portfolios, parallel forms reliability may not even be a sensible consideration.

Another issue is the rater aspect. Multiple choice tests are generally viewed as being objectively and consistently scored. Performance assessments may be scored differently depending on who does the scoring.

Performance assessments often have very few score points, such as the situation the several of the papers in this report deal with, where level scores are reported. This impacts some types of reliability estimates.

Given the arena of performance assessment, aspects of reliability need to be further examined.

Definitions of Reliability

Reliability can be conceptualized in different manners, and how it is defined and computed should influence how it is interpreted. Conceptually, test users appear to believe reliability has something to do with consistency, or getting the same 'score' twice, but often there is no distinction beyond that.

In multiple choice settings, reliability is often viewed as dealing with stability, equivalence, or both, and various methods have been derived to provide estimates of these types of reliability. Performance assessment adds the aspect of rater/scorer consistency. Factors influencing reliability values include the objectivity of the task/item/scoring, the difficulty of the task/item, the group homogeneity of the examinees/raters, speededness, number of tasks/items/raters, and the domain coverage.

Not all of these factors affect each type of reliability estimate, or influence multiple choice and performance assessments equally.

How one intends to use an assessment should determine which type of reliability estimate is of most interest. The papers in this report use different approaches to examining reliability of performance assessments. The Gao and Colton (1997) paper examines reliability from a parallel forms framework. The Wang, Kolen and Harris (1997) and Wang (1997) papers assume a psychometric model (IRT or congeneric model) in examining weighting schemes and in looking at internal consistency estimates of reliability and conditional standard errors. In contrast, the Colton, Gao and Kolen (1997) paper uses bootstrapping, and therefore does not require a strong psychometric model.

Other factors such as rater effects, whether facets are considered fixed or random in a generalizability model, whether ranking examinees or decision consistency is of more interest, how important being able to generalize to a domain is for individual examinees, which types of errors have the harshest consequences, also need to be considered in determining which reliabilities matter most in a given situation. Additionally, the interaction between validity and reliability needs to be considered. For example, it may be easier to develop comparable forms by limiting the table of specifications, but this would alter the domain that could be generalized to. Also, it may be possible to increase rater consistency by more rigidly defining scoring rubrics, but again, this might limit the generalizability.

Many reliability values are often reported for any given instrument. The purpose one has in mind for testing should color how these various values are interpreted, weighted, and used in decision making.

How to Use Reliability Values

The APA Standards (1985) emphasize the importance of identifying sources and the magnitude of measurement error, but there is not clear guidance on what to do with the information, especially in an arena such as performance assessment where consistency in scores/ lessening measurement error tends to be bought at the price of limiting/lessening validity, in terms of generalizing to the domain of interest. That is, with so few items on an instrument, increasing parallel forms reliability coefficients as a surrogate index to generalizing to the entire domain may require the constraining the domain of focus. Likewise, to increase the consistency of raters, it may be that the scoring criteria need to become more rigid, thus again limiting some of the scope of coverage.

The purpose of the rest of this paper is to sketch out some issues relating to using reliability indices (including standard errors) in the performance assessment arena.

Selecting a test

The first focus in choosing an assessment is to determine if it indeed measuress what you are trying to assess (validity), then to determine if it measures with consistency (reliability). What one is trying to measure and the uses one plans to make of the results will affect the judgment on how reliable a test needs to be. There is definitely a trade-off between reliability and validity in the performance assessment area. Having an instrument that samples from a large well defined domain may be desirable, but if each individual form of the assessment can only cover a small portion of the domain, reliability in terms of generalizing to a domain score will be severely jeopardized. However, if one is interested in a classroom level score (matrix sampling or NAEP-like), this may not be a serious constraint *if* content coverage is adequate over some reasonable number of forms. However, at an individual level, this instrument would not be adequate. Therefore, for individual level scores, it may be necessary to decrease validity in terms of constraining the domain of interest somewhat in order to obtain a more reliable estimate of an examinee's domain score. Another alternative may be to complement the performance assessment with a multiple choice measure.

There is no magical cutoff to determine if a reliability value is adequate for one's intended purpose. More is generally better than less, but a small decrease in validity may offset a larger increase in reliability. The purpose of testing needs to be considered carefully in determining how to interpret reliability estimates, and it should be recalled that the severity of consequences of measurement errors are not equal. For example, certification or admissions decisions may require a higher level of reliability than norm-referenced tests used for program evaluation or instructional effectiveness. Likewise, errors of classification may not be equally important to errors of generalizing to a domain in a given situation.

Selecting scores/forming a composite

Wang's (1997) paper discusses using reliability as a way to select weights to form a composite. This may not be an optimal way to select weights in all situations, but does give a criterion for selecting weights, given a definition of reliability. (For example, equal weights may be used when there does not appear a logical basis for unequally weighting).

Reporting scores

In performance assessment, a raw score is often reported because the way the task is scored often results in a raw score having inherent meaning, in that it is directly tied to the scoring rubric. However, there is sometimes a need to have comparable scores over time, which generally means over tasks/forms. Reliability values may be used to help select a score scale. For example, several methods of dealing with prompt raw scores were considered in deriving a score for Work Keys Listening and Writing Tests (see Wang, Kolen, & Harris, 1997). The reliability of the various scores was one aspect considered in selecting the operational method of reporting scores.

A prime consideration is that reliability be considered on both the raw scores, and on the scores that are actually reported and used. Relatively small measurement error in determining raw scores will not necessarily translate to small measurement error in derived scores based on those raw scores. This may be especially important in situations using IRT, where the responses/ratings to the tasks/items are translated to a reported score in a rather complicated fashion, or when there are a small number of scale score points.

Choosing a cut score

When cut scores are used, they should be based on content considerations, but decision consistency is also an issue. For example, setting a criterion at a level where no consistency is found will be problematic, regardless of the logical basis involved in setting it.

Comparability of forms/instruments

When one is comparing different forms or instruments, such as trying to determine if two modes of testing are interchangeable or if a less expensive test may be substituted for a more expensive version, reliability considerations may help inform the judgment. For example, when comparing two forms, the generalizibility coefficients may be one way of examining the similarities between the forms.

Choosing test length

Reliability values may be examined to determine if they appear adequate for the purposes of the assessment. The trade-offs between the length of the assessment and the validity, especially in terms of content coverage and comparability of forms, may be considered in light of logistical and fiscal concerns.

Choosing raters

Raters are an important component in obtaining performance assessment scores, and reliability indices can help inform on several decisions regarding raters. How one is conceptualizing the rater pool needs to be determined. For example, is a specific group of raters all that is of interest (such as employees at a national scoring center) , or is there a domain of raters one would like to generalize to (such as all qualified applicants who might answer an ad to become raters operationally)? Raters may have different outlooks, and different view points, experiences, etc. that they bring to the task. Are these important aspects to include? For example, should a variety of viewpoints be used in determining the quality of a piece of prose writing, or is it important that the raters have the same view point, for example, such as in judging some aspects of a liscensure test?

The comparability of raters over time with their own previous ratings, and across raters (and thus the comparability of scores) are important components of establishing trend data, or trying to chart examinee progress over time. Whether to retain a particular rater can be examined using consistency with his/her own ratings over time, and consistency with other raters, and with 'master' raters. The number of raters to employ may also be examined using reliability values, noting the expected increase in consistency for each additional rater per examinee.

An important issue that appears to be much neglected is how reliability values obtained using a national scoring center translate to local scoring; and how results from one local site generalize to others. This is directly affected by the consistency of trainers and training materials across settings, as well as the 'qualifying' measures that are used at each location.

Rater inconsistency can be due to inadequate training of raters, or inadequate specification of the scoring rubrics, or the inability of the raters to internalize the rubrics. An interesting factor of rater reliability is how it is viewed in the literature. Generally it has been found that it is possible to define rubrics so well that raters can be trained to score reliably. Currently, progress is made using computers to score written essays, demonstrating that it is indeed possible to score a well-defined task in at least some instances with computers. It is interesting, therefore, that most of the focus on use of reliability with performance assessment focuses on rater aspects, rather than on generalizing to a domain. This is unfortunate, as score reliability is generally lower than rater consistency. And increasing the number of raters is generally a less effective strategy than increasing the number of tasks or items on a test in terms of increasing reliability for score use. (See Gipps, 1994). This is especially true when the desired responses can be codified in a qualified sense--such as key words or phrases, conventions, length of response.

How much to weight/interpret score

A score that is subject to a great deal of measurement error should be interpreted more cautiously than a score that appears subject to little measurement error (assuming the interpretations are accurate with respect to validity issues). Another consequence of low reliability is not to use the scores for important decisions.

One of the purposes of reliability values are to communicate to an examinee the uncertainty in his/her score, and to alert the user of test scores regarding the replicability of the scores. Usually uncertainty is communicated using a standard error of measurement, or error bands. With some performance assessments, there may be too few points for these to be the best way to communicate information. For example, some performance assessments have taken to providing level scores, where 3-5 levels are not uncommon. In these cases, using a distribution of level scores conditional on performance to illustrate an examinee's chances of truly being at the level designated, above that level, or below that level, may all be illustrated using distributions. Distributions may be more interpretable than, say, standard errors, to both the examinee and the user of test scores. This may therefore provide information helpful in determining how likely a particular score is, and how much weight should be given it in making decisions, such as course placement.

Summary

This paper addresses the use of reliability information in choosing a score scale, forming a composite score, choosing a cut score, selecting a test, and similar situations. Making a decision by choosing the highest reliability estimate does not always appear to be the optimal decision, particularly when reliability can be assessed in different ways (e.g., rater agreement, generalizability coefficient, using a theoretical model, or bootstrapping), and when the typical reliability estimate used with performance assessments, rater agreement, may not be the most relevant, given the purpose of the assessment. Test users are encouraged to consider what definition of reliability is most meaningful, given their setting, and to make use of the reliability estimates in decision making.

References

American Psychological Association. (1985). *Standards for educational and psychological testing*. Washington, DC.

Colton, D. A., Gao, X., & Kolen, M. J. (1996). *Assessing the reliability of performance level scores using bootstrapping* . ACT Research Report 97-3. Iowa City: IA. ACT, Inc.

Gao, X. & Colton, D. A. (1996). *Evaluating measurement precision of performance assessment with multiple forms, raters, and tasks* . ACT Research Report 97-3. Iowa City: IA. ACT, Inc.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment.* Falmer Press, London.

Wang, T. (1996). *Weights that maximize reliability under a congeneric model of performance assessment.* ACT Research Report 97-3. Iowa City: IA. ACT, Inc.

Wang, T., Kolen, M. J., & Harris, D. J. (1996). *Conditional standard errors, reliability, and decision consistency of performance levels using polytomous IRT.* ACT Research Report 97-3. Iowa City: IA. ACT, Inc.

Conditional Standard Errors, Reliability and Decision Consistency of Performance Levels Using Polytomous IRT

Tianyou Wang, Michael J. Kolen, Deborah J. Harris

Abstract

This paper describes two polytomous IRT-based procedures for computing conditional standard error of measurement (CSEM) for scale scores and classification consistency indices for performance level scores. These procedures are expansions of similar procedures proposed by Kolen, Zeng and Hanson (1996) and Hanson and Brennan (1990) on different reliability indices. The expansions are in two directions. One is from dichotomous items to polytomous items and the other is from dichotomous (two-level) classification to multi-level classification. The focus of the paper is on performance assessments where the final reported scores are on a performance level scale with fewer points than traditional score scales. The procedures are applied to real test data to demonstrated their usefulness. Two polytomous IRT models were compared, and also a classical test theory based procedure for assessing CSEM was included for comparison. The results show that the procedures work reasonably well and are useful in assessing various types of reliability indices.

Conditional Standard Errors, Reliability and Decision Consistency of Performance Levels Using Polytomous IRT

Performance assessment items are usually scored on a polytomous score scale. In some testing programs (e.g., Work Keys, ACT 1995), the final reported scores are on a performance level type of scale, i.e., the examinees are classified into a finite number of levels of performance. Classifications are often based on converting raw scores to levels, because levels are relatively easy to use. In other testing programs, total raw scores are converted to reported scale scores using some linear or non-linear transformation. In either case, it is useful to obtain and report information about the conditional standard error of measurement (CSEM, conditioned at each level score or scale score), and the overall reliability. In the case of performance levels, it is also helpful to report information about classification decision consistency. To provide test users with the above information is in accordance with the recommendation by the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1985), especially Standards 2.10, 2.12, and 11.3.

Kolen, Hanson, and Brennan (1992) presented a procedure for assessing the CSEM of scale scores using a strong true-score model. In that article, they also investigated ways of using non-linear transformations from number-correct raw score to scale score to equalize the conditional standard error along the reported score scale -- a property that facilitates score interpretation. Kolen, Zeng, and Hanson (1996) presented a similar procedure for assessing the CSEM, but used item response theory (IRT) techniques. Both of these procedures were primarily developed for tests with dichotomously scored items and for scale scores. The primary purpose of this paper is to extend the procedure described in Kolen et al. (1996) to tests with polytomous items using a polytomous IRT model approach. A second purpose is to adapt the procedure to performance level scores and to discuss the similarity and difference between scale scores and level scores. A third purpose of this paper is to describe a polytomous IRT-based procedure for assessing decision consistency of performance level classification based on alternate test forms, which is also an expansion of a similar procedure by Hanson and Brennan (1990) based on the strong true score model.

Performance level scores differ from scale scores in three primary aspects. First, performance level scores usually have fewer score points than scale scores. Second, scale scores are usually transformed from the total raw scores whereas the derivation of the level scores may not necessarily be based on the total raw scores. Third, different reliability conceptions and indices might be appropriate for these two types of scores. Typically, scale scores are regarded as discrete points on a continuum. Indices such as the standard error of measurement (SEM), and parallel form reliability naturally applies to scale scores. On the other hand, level scores might be viewed as only ordered nominal categories, i.e., the numerical numbers assigned to the levels are just nominal symbols and do not have real numerical values. In this case, only classification consistency indices apply to the level scores. In some situations as in the examples in this paper, however, levels scores can also be viewed as scale scores. In this case, both SEM type of indices and classification consistency indices apply to the level scores.

In the next section, two polytomous IRT-based procedures are described. The first procedure, which can be used to assess CSEM and reliability, applies to both scale scores and performance level scores. The second procedure, which can be used to assess decision consistency, only applies to performance level scores. After the descriptions, some examples are given using some real test data to demonstrate the usefulness of these procedures.

IRT Procedure for CSEM and Reliability

The general approach for assessing the CSEM and reliability is the same as the procedure described in Kolen et al. (1996). The central task is to first obtain the probability distribution of the performance level score (or scale score) conditioned on a given θ and then compute the conditional mean and conditional standard deviation (or variance) of the scale scores or the level scores. The CSEM of the level score is the conditional standard deviation. Given a θ distribution for an examinee population, conditional means and conditional variances can be integrated over the θ distribution to obtain the overall error variance and true score variance. Reliability can thus be computed based on this information. The main difference between the present procedure and the one described in Kolen et al. (1996) lies in the step for obtaining the conditional level score distribution. In their procedure, the scale scores are converted from the total raw scores using some non-linear conversion table. As mentioned previously, the derivation of level scores might not be based directly on total raw score. In the examples of this paper using the Work Keys (ACT, 1995) tests, the conversion was originally based on a ninth order statistic from the 12 ratings given by two raters on six items. In this paper, we will describe in detail the computation procedure for level scores derived from the total raw scores and will provide some general guidelines for computing conditional standard errors for level scores that are not derived from total raw scores.

The Polytomous IRT Probability Models

Various polytomous IRT models have been developed: nominal response model (Bock, 1972), rating scale model (Andrich, 1978), graded response model (Samejima, 1969), partial credit model (Masters, 1982), and generalized partial credit model (Muraki, 1992), etc. With any of these models fitted to the polytomous test data, the probability of getting a particular response on a polytomously scored items can be computed given a θ value. In the present paper, the (generalized) partial credit model is used to fit the test data, though the polytomous IRT-based procedures described in this paper apply with any of the models just mentioned. Let U_k be the random variable for the score on item *k* with scores from 0 to m. With the generalized partial credit model, the probability of getting a particular response j is given by

$$\Pr(U_k = j|\theta) = \frac{\exp\left[\sum_{v=0}^{j} a_k(\theta - b_k + d_v)\right]}{\sum_{c=0}^{m}\exp\left[\sum_{v=0}^{c} a_k(\theta - b_k + d_v)\right]}, \qquad (1)$$

where a_k is the discrimination parameter, b_k is the difficulty parameter, and d_{kv} (v=0, 1, ... m) are the category parameters for item k.

Conditional Distribution of Raw Total Scores

Assume there are K polytomous items and let U_k be a random variable for the score on item k $(U_k = 0,1,\ldots,n_k)$. Let $\Pr(X = x|\theta)$ $(x = 0,1,\ldots,T)$ represent the conditional distribution of the raw total score $\left[X = \sum_{k=1}^{K} u_k\right]$. For dichotomous items, this distribution is a compound binomial distribution as indicated by Lord (1980). Lord and Wingersky (1984) provided a recursive algorithm for computing this distribution. For polytomous items, this distribution is a compound multinomial distribution. Hanson (1994) extended the Lord-Wingersky algorithm to the polytomous items. (The same extension was also provided by Thissen, Pommerich, Billeaud, & Williams, 1995.) This recursive algorithm is described as the following:

Let $Y_k = \sum_{j=1}^{k} U_j$ with $X = Y_K$.

For item $k = 1$,

$$\Pr(Y_1 = x|\theta) = \Pr(U_1 = x|\theta), \quad \text{for } x = 0,1,\ldots,n_1. \tag{2}$$

For item $k = 2,\ldots,K$,

$$\Pr(Y_k = x|\theta) = \sum_{u=0}^{n_k} \Pr(Y_{k-1} = x - u|\theta)\Pr(U_k = u|\theta), \quad \text{for } x = 0,1,\ldots,\sum_{j=1}^{k} n_j \tag{3}$$

$\Pr(U_k = u|\theta)$ is given by Equation 1 if a generalized partial credit model is used. The total raw score distribution is obtained after all the K items are included in this recursive procedure.

With this algorithm, we can compute the conditional distribution $\Pr(X = x|\theta)$ $(x = 0,1,\ldots,T)$, where $T = \sum_{k=1}^{K} n_k$.

Conditional Distribution of Level Scores

If the level scores are derived from the total raw score, the following procedure can be used to compute the conditional distribution of level scores. Let S symbolize the raw-to-level transformation, following the same logic as in Kolen et al. (1996), the conditional distribution of the level scores can be expressed as:

$$\Pr[S(x)=s|\theta]=\sum_{x:S(x)=s}\Pr(X=x|\theta)\ ,\quad s=1,2,\ldots,L \tag{4}$$

The mean and variance of the conditional level score distribution are:

$$\xi(\theta)=E[S(X)|\theta]=\sum_{s=1}^{L}s\Pr(S(x)=s|\theta) \tag{5}$$

$$\sigma^2[S(X)|\theta]=E\{[S(X)-\xi(\theta)]^2|\theta\}=\sum_{s=1}^{L}s^2\Pr(S(x)=s|\theta)-\xi(\theta)^2 \tag{6}$$

The conditional mean is just the conditional true level score; the square root of the conditional variance is the CSEM conditioned on θ scale. To find the CSEM conditioned on the level score, let $\eta=\xi(\theta)$ and express θ in terms of level score: $\theta=\xi^{-1}(\eta)$. Substituting this expression in Equation 6 yields the CSEM conditioned on level score η.

If the level scores are not converted directly from the raw total scores, then the conditional distribution of the level scores must be obtained by the other approaches. If the theoretical distribution of the statistic that is used to compute the level scores is known, then a theoretical approach can be used. In settings where theoretical approaches are not available, simulation techniques can be used to estimate the conditional distribution of the level scores. The steps after that are the same as those described by Equations 5 and 6. The simulation technique to obtain the conditional level score distribution is illustrated in a later part of this paper.

Average Error Variance and Reliability

Following the same logic used by Kolen et al. (1992), and Kolen et al. (1996), average error variance and reliability of the total raw scores and level scores can be computed if a θ distribution is given. To summarize their procedure, the following equations are presented here.

Let E denote error scores, $\Psi(\theta)$ denote the distribution θ, and the subscript S denote level scores. The average error variance is given by

$$\sigma^2(E) = \int_\theta \sigma^2(X|\theta)\Psi(\theta)d\theta \tag{7}$$

The marginal distribution of the raw total score is obtained by

$$\Pr(X = x) = \int_\theta \Pr(X = x|\theta)\Psi(\theta)d\theta \tag{8}$$

The observed raw score variance $\sigma^2(X)$ is the variance of this marginal score distribution. The reliability of raw total scores is given by

$$rel_{raw} = 1 - \frac{\sigma^2(E)}{\sigma^2(X)} \tag{9}$$

Similarly, average error variance for the level scores and observed score variance for level scores can be computed also by substituting $S(X)$ for X in Equations 7 and 8.

$$\sigma^2(E_s) = \int_\theta \sigma^2(S|\theta)\Psi(\theta)d\theta \tag{10}$$

$$\Pr(S = s) = \int_\theta \Pr(S = s|\theta)\Psi(\theta)d\theta \tag{11}$$

The observed level score variance $\sigma^2(S)$ is the variance of this marginal score distribution. The reliability of the level scores is then given by

$$rel_{scale} = 1 - \frac{\sigma^2(E_s)}{\sigma^2(S)} \qquad (12)$$

IRT Procedure for Decision Consistency

Decision consistency is an important type of reliability concept for educational assessments that involve classification decisions. Previous literature in this area mostly dealt with dichotomous classifications, namely mastery and non-mastery, and with dichotomous items. In recent years, both performance assessments with polytomous scoring and multiple level classification have become more popular practices. However, most of the procedures and indices developed for dichotomous classifications can be extended to multiple level classifications and to situations with polytomous items.

Usually, assessing decision consistency requires a data collection design that would require each examinee to take more than one test form. With this type of data, for a test of *L* performance levels, a *L* X *L* contingency table can be constructed based on examinees' performance levels on two test forms. Subkoviak (1984) presented some decision consistency indices which can be computed based on the contingency table. In many testing situations, however, this type of data collection design is not feasible due to time constraints or other conditions. With each examinee only taking one test form, conventional methods for assessing decision consistency do not readily apply. Subkoviak (1984) reviewed several alternative procedures using stronger statistical assumptions for obtaining the contingency table based on data from a single form. For example, Huynh (1976) proposed a complicated method based on Keats and Lord's (1962) beta-binomial model. Subkoviak (1976) proposed a simpler method which uses the binomial distribution for the approximation. Hanson and Brennan (1990) proposed a method using a strong true score model. In their procedure, the strong true score model is used to compute the conditional contingency table and the conditional contingency table was integrated over a distribution of the true score. Their methods might be extended to polytomous classification case. IRT-based techniques, however, can also be used to obtain the contingency table based on data from a single form. In the next sections, we will first discuss some important decision consistency indices and some procedures for calculating these indices and then we will describe in detail an IRT-based procedure to assess decision consistency for multiple level classifications. This procedure is a natural extension of the Hanson and Brennan (1990) procedure except it uses IRT techniques instead of the strong true score model.

Decision Consistency Indices

Subkoviak (1984) reviewed some decision consistency indices and procedures to compute them. Based on his review, two important indices are p_0 and κ. p_0 is the proportion of consistent classification based on two parallel forms and κ is the proportion of consistent classification adjusted for chance. These two indices are computed based on a classification outcome table (we will refer to it as the contingency table) as illustrated by the following hypothetical example of 3-level classification:

		Form 2			
		Level 1	*Level 2*	*Level 3*	*Low total*
Form 1	*Level 1*	*0.12*	*0.11*	*0.07*	*0.30*
	Level 2	*0.09*	*0.23*	*0.08*	*0.40*
	Level 3	*0.03*	*0.10*	*0.17*	*0.30*
Column Total		*0.24*	*0.44*	*0.32*	*1.00*

With this table, consistent proportion $p_0 = \sum_{k=1}^{L} \hat{p}_{kk} = 0.12 + 0.23 + 0.17 = 0.52$. In order to compute κ, we must first compute p_c which is the consistent proportion due to pure chance, that is, if the two forms are independent and the entries of the table would be the product of the corresponding row and column totals. So $p_c = \sum_{k=1}^{L} \hat{p}_{k.}\hat{p}_{.k} = 0.30{*}0.24 + 0.40{*}0.44 + 0.30{*}0.32 = 0.346$. And

$$\kappa = \frac{p_0 - p_c}{1 - p_c} = \frac{0.52 - 0.346}{1 - 0.346} = 0.266 \ . \tag{13}$$

It is observed that with multiple levels the classification indices usually appear to be lower than classification with dichotomous levels. Taking the above example as an illustration, if level 1 and level 2 are collapsed into a single level, then we have a dichotomous classification problem. Following the same computational procedure with the collapsed 2 X 2 contingency table, it is found that p_0=.72, and κ=.346.

IRT Procedure for Obtaining the Contingency Table

Earlier in this paper, procedures for computing the conditional distribution of the level scores based on the item parameters and the classification criterion were presented. Here it is assumed that the conditional distribution of the level scores are known. By the assumption of local independence, the responses to items in two parallel forms are independent conditioned on a single θ value, thus the classifications based on the two forms are also independent given that θ. Based on this result, a conditional contingency table can be computed by multiplying the corresponding conditional level score probabilities. Let $\left[A_{ij} \mid \theta\right]$ denote the entry *i, j* of the conditional contingency table, it is expressed as the following:

$$\left[A_{ij} \mid \theta\right] = \Pr\left(S_1 = i \mid \theta\right) \Pr\left(S_2 = j \mid \theta\right) \ , \ i = 1,2,\ldots,L, \ \ j = 1,2,\ldots,L \ . \tag{14}$$

Given a θ population distribution $\Psi(\theta)$, the overall contingency table can be computed by integrating the conditional contingency table over $\Psi(\theta)$ using numerical integration.

$$\left[A_{ij}\right] = \int_{\theta} \left[A_{ij} \mid \theta\right] \Psi(\theta) d\theta \ , \ i = 1,2,\ldots,L, \ \ j = 1,2,\ldots,L \ . \tag{15}$$

After the contingency table is obtained, decision consistency indices p_0 and κ can be computed using the procedure illustrated in the hypothetical example in the previous section.

Examples with Real Test Data

Tests and Test Scores

In this paper, Work Keys (ACT, 1995) test data are used to illustrate the computations and usefulness of the previously described IRT-based procedures. The Work Keys Listening and Writing tests each consist of 6 items. The response to each item is rated on a 0 to 5 rating scale by two independent raters. When the two ratings differ by more than one point, a third "expert" rater is used. In this case, the third rater's ratings replace the ratings of each of the first two raters. Each examinee has 12 ratings.

The analyses here involved two types of performance level scores. The original performance levels correspond to the ninth order statistic of the 12 ratings. That is, the 12 ratings are sorted from highest the to the lowest, and the ninth from the highest score is the level score assigned to this examinee. This process was followed so that 75% of the 12 ratings earned by the examinee would be at or above the reported level. In the scale development process for the Work Keys tests, some alternative procedures for assigning level scores have been considered, one of which is to assign level score according to the rounded mean of the 12 ratings. For convenience, we will call the former the old level scores and the latter new level scores. It is also desirable to compare the psychometric properties of this new level score assignment procedure.

The analyses reported here used data for three forms (10, 11, and 12) of the Work Keys Writing tests. (Because of the limited space of the paper and because of a calibration problem with the Listening test using the PARSCALE program, only the results for the Writing test is presented in this paper. Interested readers can refer to Wang, Kolen & Harris, 1996, for the results for the Listening test calibrated using the FACETS program.) The sample sizes for these three forms are 7097, 2035, and 1793, respectively.

Method and Analysis

The primary purpose of this example is to illustrate the polytomous-IRT based procedures for estimating CSEM, reliability and decision consistency indices described in this paper. A second purpose is to compare the partial credit model which has a fixed discrimination parameter across items with the generalized partial credit model which has varying discrimination parameters across items. A third purpose is to compare the IRT based procedure with a classical test theory-based procedure described in Feldt and Qualls (1996) for estimating CSEM.

A partial credit model (Masters, 1982) was fit to the response data using the FACETS (Linacre, 1989) computer program. A generalized partial credit model (Muraki, 1992) was fit to the same data using the PARSCALE (Muraki & Bock, 1993) program. Each item originally has 6 score categories (0 to 5). The sum of two ratings results in 11 score categories for each of the six items. Because the sample size for Form 10 is too large for FACETS, only half of the data set (every other examinee) was used in the calibration.

Because it is cumbersome to derive the conditional distribution of the ninth order statistic based on a probability model, a simulation technique was used to find the conditional distribution of the old (ninth order statistics-based) level score. The steps after that are the same as those for the new level score.

Computational Steps for the Polytomous IRT-Based Procedure for CSEM

For the new level scores, the computation follows these steps:

(1) Conditioned on a certain point on the θ scale, (a) the raw score distributions were estimated using the extended Lord-Wingersky algorithm in Equations 2 and 3, (b) the level score distributions were estimated using Equation 4 with the raw score-to-level score conversion table, (c) the conditional expected (true) level score and error variance were computed using Equation 5 and 6, (d) the conditional 6x6 contingency tables were computed using Equation 14.

(2) Using an empirical θ distribution, the following overall indices were computed using numerical integration: (a) error level score variance, observed level score variance and true level score variance, and reliability (using Equations 10, 11, and 12), (b) the overall contingency table (Equation 15) and classification indices (p_0 and κ) (Equation 13), (c) the marginal distribution of the level scores (Equation 11). To obtain the empirical θ distribution, the θ estimates for examinees from the FACETS output are used whereas a directly estimated θ distribution is output from the PARSCALE program.

For the old level scores, the computation follows these steps:

1) Conditioned on a quadrature point on the θ scale, simulate responses to each of the six items for 200 simulees with the same theta. Each response, which ranges from 0 to 10, were broken into two ratings which range from 0 to 5 based on the rule that the two ratings can not differ more than one point. For instance, a score of 9 was broken into 4 and 5, and a score of 8 was broken into 4 and 4, etc. The ninth order statistic was used as the old level score. Based on these simulated data, (a) the level score distribution, (b) the conditional mean (true) level score and error variance (Equations 5 and 6), (c) the conditional 6x6 contingency table (Equation 14) were computed.

(2) Using an empirical θ distribution based on the θ estimates from the FACETS output, the following overall indices were computed using numerical integration: (a) error level score variance, observed level score variance and true level score variance, and reliability (using Equations 10, 11, and 12), (b) the overall contingency table (Equation 15) and classification indices (p_0 and κ) (Equation 13), (c) the marginal distribution of the level scores (Equation 11).

The Feldt/Qualls Procedure for Estimating CSEM

Feldt and Qualls (1996) proposed a procedure for computing the CSEM which is a modification of the Thorndike's (1951) procedure. This procedure assumes that the test consists of *d* essentially tau-equivalent parts and uses the square term of the parts difference scores to estimate the error variance. (For details see their paper.) In the present study, we assume the six items are six essentially tau-equivalent parts and mainly use Equation 7 in the Feldt and Qualls (1996) paper. Although the assumption of essential tau-equivalency may be violated in our case, it was considered useful to use this procedure to provide some comparisons.

Results

Model fit was partially checked by comparing the expected score distribution based on the model and the actual score distribution based on the test data. The fitted total score distribution was computed based on Equation 11. Figures 1 plots the fitted and observed total score distributions. It was found that the fitted distributions were close to the observed score distributions both for the FACETS and PARSCALE models, suggesting that both the partial credit model and the generalized partial credit models fit reasonably well. Note, however, that for the FACETS model the upper tail of the fitted distribution is somewhat higher than the tail for the observed distribution. This might have resulted from using the examinee ability estimates in the integration process. These higher tails are consistent with a similar finding discussed by Han, Kolen and Pohlmann (1997) for multiple choice tests.

Figure 2 contains plots of the conditional expected (true) level scores for the old and new levels using both FACETS and PARSCALE models. These plots consistently show that the new level scores are easier than the old level scores, particularly at low levels. This result is not surprising because the mean score corresponds to the 6th or 7th order statistic, which is easier than the 9th order statistic. The plots of the expected levels are quite close for the two models.

Tables 1 and 2 contains the marginal distributions of the old and new level scores for the FACETS and PARSCALE models. Comparisons between the old and new level scores are consistent with the trends shown in Figure 2. For the old level scores, the estimated marginal level score distributions are flatter than the observed level score distribution. For the new level scores, the estimated marginal level score distributions are quite close to the observed level score distributions, particularly with the PARSCALE model. These results suggest that the polytomous

IRT models fit the data better at aggregate score level (from which the new level scores are derived) than at individual item level (from which the old level scores are derived).

The conditional standard errors (CSEM) of the old and new level scores are presented in Figures 3 and 4. Figure 3 plots CSEM along the θ scale whereas Figure 4 plots CSEM along the level score scale. The conditional level scores in Figure 4 are the expected level scores conditioned on θ and can be regarded as the true level scores according to the usual definition. Thus, fractional true level scores are possible whereas in reality fractional observed level scores are not possible. These plots show that CSEM for the old and new levels have quite different patterns. The old level scores have big CSEM around level one. Generally, the old level scores have larger CSEM than the new level scores. Generally, the CSEM of the new level scores bump at each level, with the mode between two adjacent level scores. The bumps resulted from the rounding in deriving the level scores. In between two adjacent level scores, the rounding will result in larger error than around each of the levels. That is, conditioned at a true level score of, say, 2.5, examinees may receive level scores of 2 or 3, thus the variance for this examinee group is much larger than the group with a true level score of 2 or 3. The bumps for the old level score do not have a clear and consistent pattern and is more difficult to explain. In general, the CSEM plots are similar for the PARSCALE and FACETS models.

The CSEM computed for the new level scores based on the Feldt and Qualls procedure are presented in Table 3. Because the conditional variable level scores are integer points, they cannot be plotted as in Figure 4. Overall, these estimates are close to the IRT-based CSEM estimates conditioned at those exact level points where there are minimal rounding errors for the IRT-based estimates. This happens because the Feldt and Qualls procedure did not take rounding error into consideration. The CSEM estimates based on Feldt and Qualls procedure decrease as level scores go from low to high. This trend can also be observed from Figure 4 for those exact level points. However, the bumpy modes in Figure 4 stay almost always constant, an interesting result not readily interpretable.

The classification consistency and reliability indices for the two models are summarized in Tables 4 and 5. Again, these results clearly suggest that the new level scores have higher reliability and classification consistency than the old level scores. This result is consistent with the findings for the CSEM. It is interesting to notice that the FACETS-based reliability and classification consistency are both slightly higher than the PARSCALE-based indices. But because we do not know the true value of these indices, it is difficult to judge which model gives more accurate estimates. Compared with the overall error variance based on the Feldt and Qualls procedure, the polytomous IRT-based overall error variance are slightly higher. This difference may reflect the fact that the IRT-based procedure can take into account the error caused by rounding.

Discussion and Conclusions

This paper described two polytomous IRT-based procedures for computing CSEM for scale scores and classification consistency indices for performance level scores. The former is a natural extension of a dichotomous IRT-based procedure by Kolen, Zeng and Hanson (1996); the latter is an expansion of a strong true score model-based procedure by Hanson and Brennan (1990) in two directions: from dichotomous items to polytomous items and from dichotomous (two-level) classification to multi-level classification. The focus of the paper is on performance assessments which normally use polytomous scoring. In particular, the procedures apply to those performance assessments where the final reported scores are on a performance level scale with fewer points than traditional score scales. Because the scoring process involves classification, classification decision consistency type of reliability indices are also relevant in addition to more conventional reliability indices.

The application of these two procedures to the Work Keys Writing assessment seems to indicate that they work reasonably well. The results demonstrate that these procedures can be used to assess the various aspects of psychometric properties of an assessment with polytomously scored items, particularly the CSEM for scale scores and classification consistency for performance level scores. The analyses also examined one scoring procedure which is not based on the total score but based on a ninth-order statistic and compared it with a new scoring procedure which is based on the total score. The results of the analyses was instrumental in the final adoption of the new scoring procedure in the Work Keys assessment program.

The polytomous IRT-based procedures proposed in this paper apply with different types of polytomous IRT models. There are two general categories of polytomous models that apply to the type of test data discussed in this paper: the graded response models and the (generalized) partial credit models. The analyses in this paper included only one of these categories even though it is expected that the procedures should work equally well with the other category of models. In particular, we applied and compared the partial credit model with FACETS and the generalized partial credit model with PARSCALE. This comparison is analogous to the Rasch model versus the two-parameter IRT models for the dichotomous items. The results indicate the two models yield slightly different results with the PARSCALE model producing marginally better results based on the criterion of the observed marginal level score distribution. Overall, the FACETS model also seems to produce reasonably accurate estimates. The comparison between the IRT -based results and the Feldt and Qualls procedure on CSEM also gave some interesting results, particularly the ability of the IRT-based procedure to take into account the error due to rounding.

In a related study by Colton, Gao and Kolen (1997) on the same Work Keys data, the bootstrapping procedure they used produced for Form 10 error variance estimates .1922 for the old level scores and .1177 for the new level scores. These error variance estimates are remarkably close to the FACETS-based estimates which are .1902 for the old level scores and .1190 for the new level scores. These results are also close to the PARSCALE-based estimates which are .2050 and .1367 respectively. Considering that these procedures used totally different methodologies, the similarity of the results provides evidence of the accuracy of the polytomous IRT-based procedures.

References

ACT (1995). *Work Keys Assessments.* Iowa City: ACT.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Colton, D. A., Gao, X., & Kolen, M. J. (1997). Assessing the reliability of performance level scores using Bootstrapping. *ACT Research Report Series, 97-3.* Iowa City, IA: ACT.

Feldt, L. S., & Qualls, A. L. (1996). Estimation of measurement error variance at specific score levels. *Journal of Educational Measurement, 33,* 141-156.

Han, T., Kolen, M. J., & Pohlmann J. (1997). A comparison among IRT true- and observed-score equatings and traditional equipercentile equating. *Applied Measurement in Education, 10,* 105-121.

Hanson, B. A. (1994). *An extension of the Lord-Wingersky algorithm to the polytomous items.* Unpublished research note.

Hanson, B. A. & Brennan, R. L. (1990). An investigation of classification consistency indices estimated under alternative strong true score models. *Journal of Educational Measurement, 27,* 345-359.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13,* 253-264.

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for the mental test scores. *Psychometrika, 27,* 59-72.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement, 29,* 285-307.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33,* 129-140.

Linacre, J. M. & Wright, B. D. (1993). *FACETS.* MESA Press: Chicago.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement,* 8, 452-461.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a mastery test. *Journal of Educational Measurement, 13,* 265-276.

Subkoviak, M. J. (1984). Estimating the reliability of mastery-nonmastery classifications. In Berk, R. A. (Ed.). *A guide to criterion-referenced test construction*. Baltimore, MD: The John Hopkins University Press.

Thissen, D., Pommerich, M. Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*, 39-49.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560-620). Washington, DC: American Council on Educational.

Wang, T., Kolen, M. J., & Harris, D. J. (1996). *Conditional Standard Errors, Reliability and Decision Consistency of Performance Levels Using Polytomous IRT.* Paper presented at the Annual Meeting of the American Educational Research Association, New York, April.

Table 1. FACETS based marginal distribution for the Writing test.

		Marginal Distribution			
		Old Level		New Level	
Form	Level	Estimated	Observed	Estimated	Observed
10	0	0.0343	0.030	0.0054	0.003
	1	0.0098	0.014	0.0245	0.018
	2	0.2450	0.245	0.1730	0.161
	3	0.4295	0.538	0.4249	0.458
	4	0.2808	0.172	0.3675	0.353
	5	0.0006	0.001	0.0047	0.007
11	0	0.0443	0.039	0.0069	0.001
	1	0.0143	0.025	0.0362	0.034
	2	0.2921	0.266	0.1930	0.179
	3	0.3742	0.495	0.3951	0.422
	4	0.2686	0.170	0.3501	0.344
	5	0.0065	0.006	0.0188	0.020
12	0	0.0912	0.091	0.0103	0.005
	1	0.0193	0.027	0.0617	0.049
	2	0.3561	0.328	0.2513	0.240
	3	0.3439	0.461	0.4319	0.467
	4	0.1886	0.093	0.2417	0.235
	5	0.0008	0.001	0.0031	0.004

Table 2. PARSCALE based marginal distribution for the Writing test.

		Marginal Distribution			
		Old Level		New Level	
Form	Level	Estimated	Observed	Estimated	Observed
10	0	0.0301	0.030	0.0036	0.003
	1	0.0089	0.014	0.0221	0.018
	2	0.2608	0.245	0.1650	0.161
	3	0.4241	0.538	0.4504	0.458
	4	0.2745	0.172	0.3539	0.353
	5	0.0016	0.001	0.0049	0.007
11	0	0.0406	0.039	0.0040	0.001
	1	0.0127	0.025	0.0328	0.034
	2	0.3037	0.266	0.1878	0.179
	3	0.3818	0.495	0.4121	0.422
	4	0.2560	0.170	0.3487	0.344
	5	0.0053	0.006	0.0146	0.020
12	0	0.0926	0.091	0.0078	0.005
	1	0.0193	0.027	0.0625	0.049
	2	0.3672	0.328	0.2351	0.240
	3	0.3588	0.461	0.4736	0.467
	4	0.1617	0.093	0.2195	0.235
	5	0.0004	0.001	0.0015	0.004

Table 3. The standard error of measurement for the level scores for the Writing test from the Feldt and Qualls procedure

Form\Level	0	1	2	3	4	5	Overall
10	0.2058	0.3761	0.2421	0.2215	0.1957	0.1526	0.2186
11	0.2541	0.3632	0.2533	0.2520	0.2209	0.1664	0.2439
12	0.2771	0.4451	0.3055	0.2518	0.2243	0.1534	0.2691

Table 4. FACETS based classification consistency and reliability indices for the Writing test.

Form	Old Level			New Level		
	po	pc	kappa	po	pc	kappa
10	0.6945	0.3247	0.5476	0.7638	0.3461	0.6387
11	0.6484	0.2997	0.4980	0.7285	0.3176	0.6021
12	0.6196	0.2894	0.4648	0.6959	0.3120	0.5579
	var(T)	var(E)	reliability	var(T)	var(E)	reliability
10	0.6793	0.1902	0.7812	0.5677	0.1190	0.8267
11	0.7897	0.2244	0.7787	0.6771	0.1372	0.8315
12	0.9226	0.2963	0.7569	0.6635	0.1566	0.8090

Table 5. PARSCALE based classification consistency and reliability indices for the Writing test.

Form	Old Level			New Level		
	po	pc	kappa	po	pc	kappa
10	0.6643	0.3242	0.5032	0.7292	0.3559	0.5796
11	0.6290	0.3054	0.4659	0.7007	0.3280	0.5547
12	0.6159	0.2987	0.4523	0.6655	0.3317	0.4995
	var(T)	var(E)	reliability	var(T)	var(E)	reliability
10	0.6358	0.2053	0.7559	0.5024	0.1367	0.7861
11	0.7279	0.2359	0.7552	0.5972	0.1522	0.7969
12	0.8948	0.2761	0.7642	0.5823	0.1727	0.7713

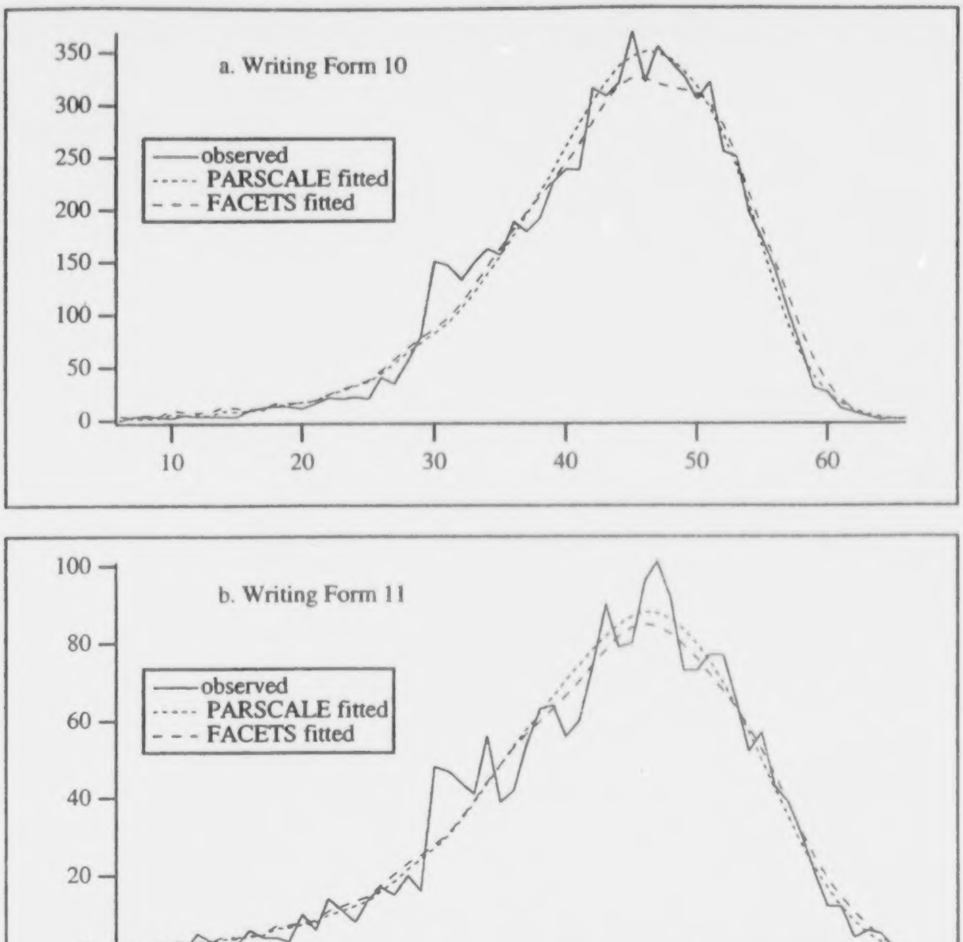

Figure 1. The fitted and observed score distributions for the Writing test.

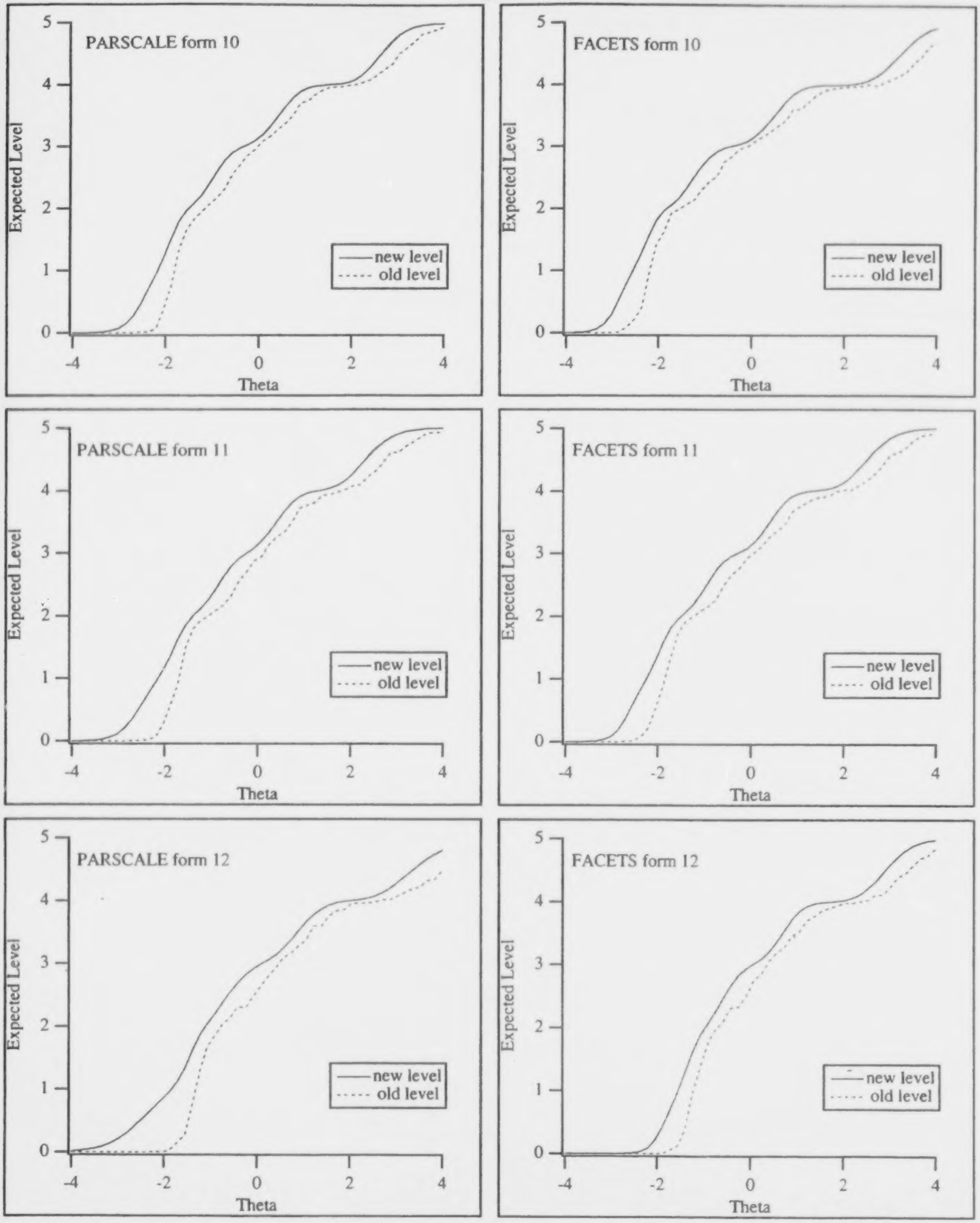

Figure 2. The conditional expected (true) level scores for old and new levels.

BEST COPY AVAILABLE

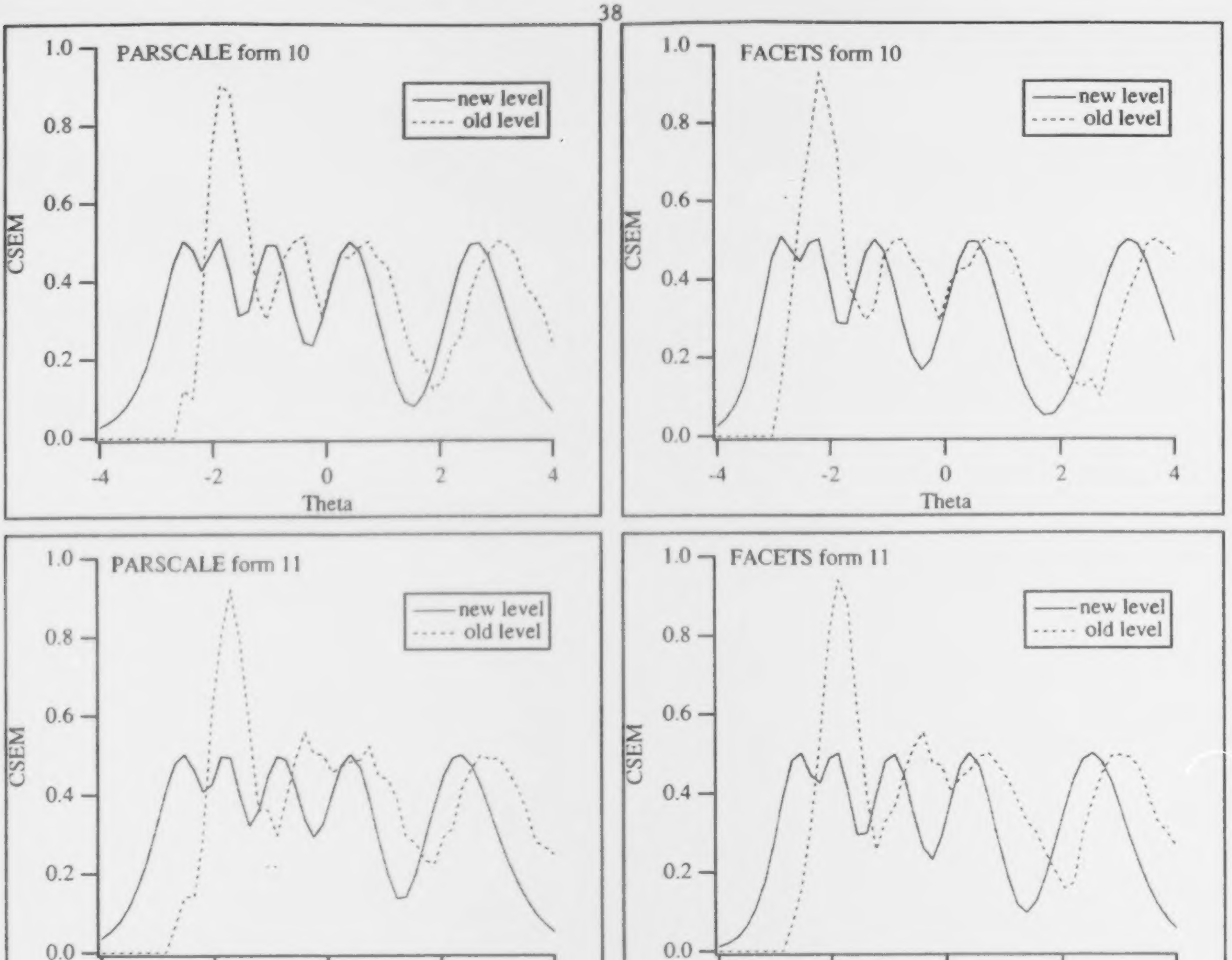

Figure 3. The conditional standard error (CSEM) for old and new levels conditioned on theta.

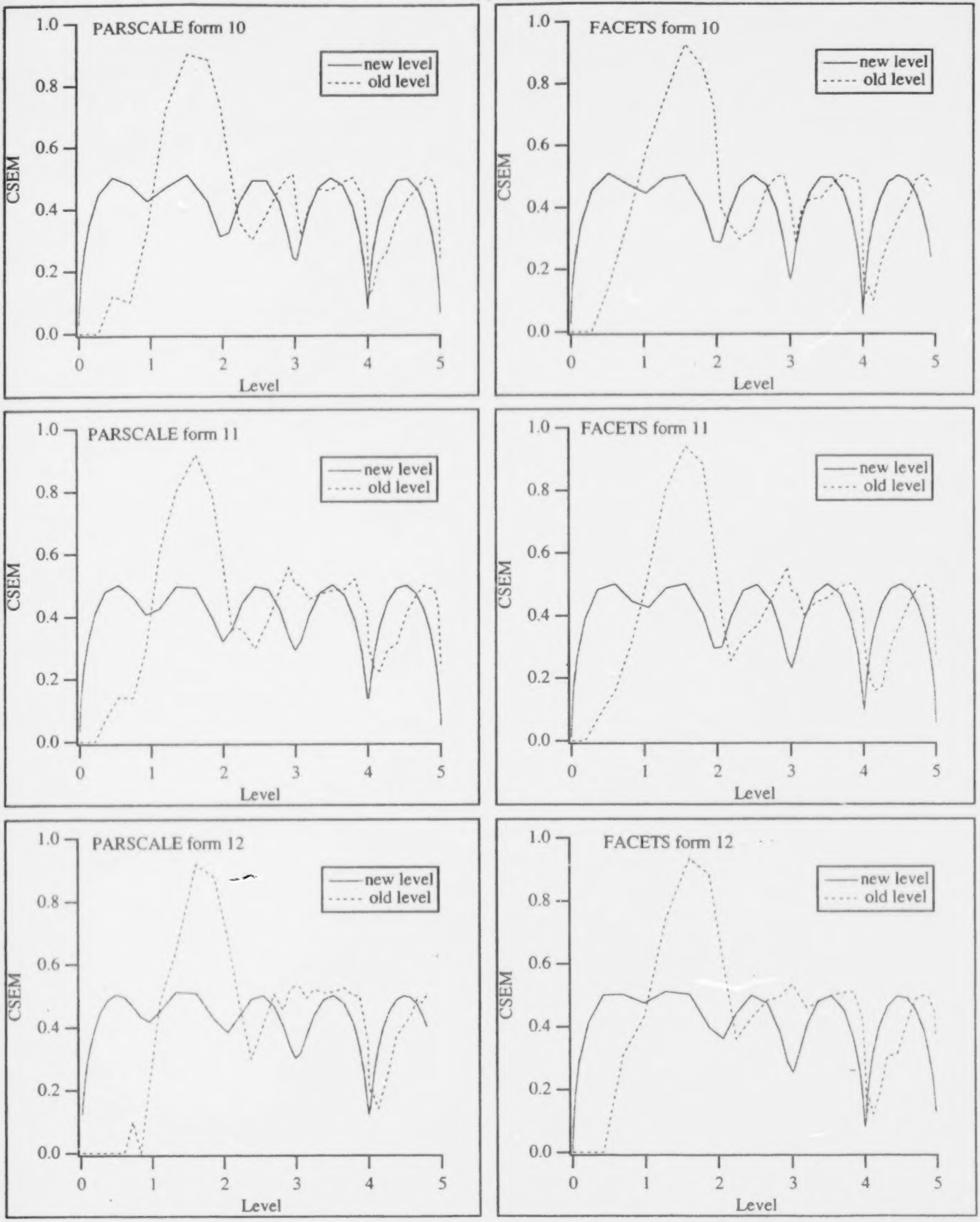

Figure 4. The conditional standard error (CSEM) for old and new levels conditioned on level.

BEST COPY AVAILABLE

Assessing the Reliability of Performance Level Scores Using Bootstrapping

Dean A. Colton, Xiaohong Gao, Michael J. Kolen

Abstract

This paper describes a bootstrap procedure for estimating the error variance and reliability of performance test scores. The bootstrap procedure is used in conjunction with generalizability analyses to produce estimated variance components, measurement error variances, and reliabilities for two types of performance scores using data from a large scale performance test that measures both listening and writing skills. The first type of score was simply the rounded average of the performance ratings. The second type of score was a performance level score related to the difficulty and complexity of the items as assembled in test development. Results on the two tests and two types of scores are reported, and the described methods are suggested for use with other performance measures.

Assessing the Reliability of Performance Level Scores Using Bootstrapping

Total raw scores for performance assessments typically are calculated by summing raw scores over raters and items. These raw scores sometimes are transformed to integer-value proficiency level scores. Although the reliability of raw scores might be readily estimated by generalizability theory (Brennan, 1993) when the sum of the scores is used, there does not appear to be a straightforward way to use generalizability theory to find reliability of scale scores that are not linear transformations of raw scores. In the present paper, the bootstrap resampling procedure (Efron & Tibshirani, 1993) is used to estimate conditional standard errors of measurement and reliability for performance level scores.

Data

The data for this study were from 7097 examinees who took Form 10 of the Work Keys Listening and Writing assessment. The Listening and Writing assessment contains six prompts (tasks). Examinees are asked to listen to six audio-taped prompts ranging from easy and short to difficult and long. After each prompt, they are told to construct a written summary about the prompt. The written responses were scored separately for Listening and Writing by two different pairs of raters. If the ratings differ by more than one point, a third "expert" rater is used. The rating of this third rater replaces the ratings of each of the first two raters. Each rating ranges from 0 to 5. For Listening or Writing, each examinee receives a total of 12 ratings (6 prompts x 2 raters).

Level Scores are reported as indicators of examinees' Listening and Writing performance. Each of the ratings in the 0 to 5 range is intended to represent the proficiency level of the examinee's response. For example, a rating of 3 is intended to indicate that the response is at Level 3. To be conservative, it was decided by Work Keys development staff that the Level Score reported to the examinee should be one at which 75% of the 12 ratings are at or above that rating. To find this Level Score, the 12 ratings are ranked from highest to lowest. The Level Score reported to the examinee is the 9th from the highest, which we refer to here as the *9th order statistic*. For example, an examinee with ratings 5, 5, 4 ,4 ,4 ,4 ,4, 4, 4, 3, 3, 3 would receive a Level Score of 4. An examinee with ratings 5, 5, 4 ,4 ,4 ,4 ,4, 4, 3, 3, 3, 3 would receive a Level Score of 3.

Because of concerns about the unreliability of Level Scores, an alternate procedure based on the rounded average was used to create *Rounded-Average Level Scores*. Each examinee's 12 ratings were summed to get a total score ranging from 0 to 60. The total score was then divided by 12 and the unrounded average score was rounded up at .5 to obtain an integer value ranging from 0 to 5. For example, a total score of 30 was averaged to 2.50 and was then rounded up to 3.

Analyses

Bootstrap procedures were used to estimate conditional standard errors of measurement at each level for both the Level Scores and the Rounded-Average Level Scores. In addition, reliabilities for both types of scores were calculated.

Bootstrap

The bootstrap procedure was implemented separately for Listening and Writing for each examinee as follows.

1. Generate a random integer from 1 to 6, and refer to this integer as i. For each examinee, select the observed Rater 1 and Rater 2 ratings on prompt i.
2. Repeat step 1, 6 times. At the conclusion of step 2, for each examinee we have 12 ratings based on selecting the prompts, with replacement.
3. For each examinee, calculate the Level Score and Rounded-Average Level Score from the 12 ratings assembled in Step 2.
4. Repeat steps 1 through 3 $n_b = 500$ times.

Following these procedures produced 500 bootstrap Level Scores and 500 bootstrap Rounded-Average Level Scores for each of the 7097 examinees.

Conditional Standard Errors of Measurement and Reliability

For examinee, p, the absolute standard error of measurement was calculated as follows:

$$\hat{\sigma}(\Delta_p) = \hat{\sigma}(X_{pB}) = \sqrt{\frac{\sum_{b=1}^{n_b} X_{pb}^2 - (\sum_{b=1}^{n_b} X_{pb})^2 / n_b}{n_b - 1}}, \qquad (1)$$

where X_{pb} is the Level Score or Rounded-Average Level Score and the summations in Equation 1 are over the $n_b = 500$ bootstrap replications. Brennan (1996) proved that the absolute standard error of measurement is the square root of the variance of a distribution of means.

Separately for each type of level score, the examinees were then assigned to six groups according to their mean score using the bootstrap data. That is, true Level Score was defined as the mean Level Score over the 500 replications, and true Rounded-Average Level Score was defined as the mean Rounded-Average Level Score over the 500 replications. In this study, the average standard errors for each level (l) were computed using the following equation:

$$\hat{\sigma}_l(\Delta) = \sqrt{\frac{1}{n_{p:l}} \sum_{p=1}^{n_{p:l}} \hat{\sigma}^2(\Delta_p)}, \qquad (2)$$

where the summation is over persons originally classified at Level l.

To find reliability coefficients, the 7089 person by 500 bootstrap sample matrix of Level Scores was treated as a person (p) by form (b) generalizability analysis and analyzed using GENOVA (Crick & Brennan, 1982). Using generalizability theory notation, the average, over examinees, of the absolute error variance, which is the square of the expression in Equation 1, can be expressed as $\hat{\sigma}^2(\Delta) = \hat{\sigma}^2(B) + \hat{\sigma}^2(pb)$, where $\hat{\sigma}^2(B)$ is the variance of form means over bootstrap replications and $\hat{\sigma}^2(pb)$ is the combined person by form interaction and residual variance. Also, the average, over examinees, relative error variance from generalizability theory can be expressed as $\hat{\sigma}^2(\delta) = \hat{\sigma}^2(pb)$. Generalizability, $E\hat{\rho}^2$, and dependability, $\hat{\Phi}$, coefficients can also be estimated for each level using the following equations:

$$E\hat{\rho}^2 = \frac{\hat{\sigma}^2(p)}{\hat{\sigma}^2(p) + \hat{\sigma}^2(\delta)}, \text{ and} \tag{3}$$

$$\hat{\Phi} = \frac{\hat{\sigma}^2(p)}{\hat{\sigma}^2(p) + \hat{\sigma}^2(\Delta)}, \tag{4}$$

where $\hat{\sigma}^2(p)$ is person variance.

Finally, to find the relative conditional standard errors, the variability due to form differences was subtracted from the error variance based on the absolute standard errors in Equation 2 as follows:

$$\hat{\sigma}_l(\delta) = \sqrt{\hat{\sigma}_l^2(\Delta) - \hat{\sigma}_l^2(B)}. \tag{5}$$

Results

The average error variances, reliabilities, and variance components are shown in Table 1. As expected, the relative error variances are smaller than the absolute error variances, and the relative generalizability coefficient is larger than the absolute generalizability coefficient. The Writing test is more reliable than the Listening test. The Rounded-Average Level Scores are more reliable than the Level Scores.

The bootstrap procedure was conducted twice for each performance test and the estimated absolute error variances were compared. For both the Level Scores and the Rounded-Average Level Scores, the absolute error variance values in the replication analysis were very close to the values obtained in the first bootstrap analysis. For the Listening test, the two estimates of absolute error variance for the Rounded-Average Scores differed in the third decimal place, and the estimates for the Level Scores differed in the second decimal place. For the Writing test, the two estimates for the Rounded-Average Scores differed in the fourth decimal place, and the estimates for the Level Scores differed in the third decimal place. Even though the bootstrap procedure was carried out by sampling from only six prompts, the estimates of absolute error variance appeared to be fairly stable.

Insert Table 1 about here

Figures 1 through 4 were constructed to display the relationship between conditional standard errors and level scores. The horizontal axis in these figures is the mean (for each examinee) level score over the 500 bootstrap replications. The vertical axis is the standard deviation of the examinee's level scores over the 500 bootstrap replications as calculated using Equation 1. One finding that is clear from these figures is that there is much less variability of the estimated standard errors for the Rounded-Average Level Scores than for the Level Scores. Also, the estimated standard errors for the Rounded-Average Level Scores tend to be lower than those for the Level Scores. There is some spread of estimated standard errors at all points on the vertical axis, and there is a tendency for the estimated standard errors to be somewhat larger at middle scores than at the more extreme scores. Some examinees had estimated standard errors of zero. (Note that when examinees have 12 identical ratings, the standard errors estimated using the bootstrap necessarily are zero.)

Insert Figures 1 through 4 about here

Mean standard errors and error variances conditional on level score as calculated using Equation 2 are given in Table 2. The standard errors differ across levels, with the largest standard errors tending to occur for examinees receiving a level score of 1. Also, the conditional standard errors for the Rounded-Average Level Scores tend to be smaller than those for the Level Scores.

Insert Table 2 about here

Discussion and Conclusions

The findings presented indicated that the Rounded-Average Level Scores tended to be more reliable than the Level Scores. This finding led the Work Keys program to reconsider the use of the Level Scores for new Level Scores that were more reliable. The findings also suggest that conditional standard errors differ across levels. This difference should be used when interpreting scores.

It should be noted that the item sampling procedure used here did not simulate item sampling as done in construction of operational forms. In the procedure used here, the sampling of items could result in form to form differences in difficulty, since items were sampled with replacement.

The methodology presented here can prove useful in situations in which ratings are nonlinearly transformed to level scores. Because the use of proficiency levels has become pervasive with performance assessments, the reliabilities and conditional standard errors of the proficiency levels need to be estimated. The methods presented here can be used to estimate these quantities.

References

Brennan, R. L. (1992). *Elements of generalizability theory.* Iowa City, IA: American College Testing.

Brennan, R. L. (1996). *Conditional standard errors of measurement in generalizability theory.* (Iowa Testing Programs Occasional Papers, No. 40).Iowa City, IA: The University of Iowa.

Crick, J. E. & Brennan, R. L. (1982). *GENOVA: A generalized analysis of variance system (FORTRAN IV computer program and manual).* Dorchester, MA: Computer Facilities, University of Massachusetts at Boston.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap* (Monographs on Statistics and Applied Probability 57). New York: Chapman & Hall.

Table 1

Results of Generalizability Analysis of Bootstrap Level Scores

Source	Listening		Writing	
	Level Score	Rounded Average Level Score	Level Score	Rounded Average Level Score
Person: $\hat{\sigma}^2(p)$	0.32803	0.31259	0.62569	0.54281
Form: $\hat{\sigma}^2(b)$	0.02451	0.02892	0.01182	0.01016
Person x Form: $\hat{\sigma}^2(pb)$	0.21548	0.15981	0.19220	0.11773
Relative Error: $\hat{\sigma}^2(\delta)$	0.21548	0.15981	0.19220	0.11773
Absolute Error: $\hat{\sigma}^2(\Delta)$	0.23999	0.18873	0.20402	0.12789
Relative G Coefficient: $E\hat{\rho}^2$	0.60354	0.66171	0.76501	0.82177
Absolute G Coefficient: Φ	0.57750	0.62353	0.75411	0.80932

Table 2

Conditional Standard Errors of Measurement

Listening

Level	Level Scores			Rounded Average Level Scores		
	Number of Examinees	$\hat{\sigma}_l(\Delta)$	$\hat{\sigma}_l(\delta)$	Number of Examinees	$\hat{\sigma}_l(\Delta)$	$\hat{\sigma}_l(\delta)$
0	129	0.50418	0.47926	23	0.33003	0.28284
1	373	0.74375	0.72708	134	0.51495	0.48606
2	4431	0.46049	0.43306	1753	0.44507	0.41130
3	2100	0.49075	0.46511	4647	0.42352	0.38788
4	64	0.54093	0.51777	535	0.47291	0.44127
5				5	0.43349	0.39874

Writing

Level	Level Scores			Rounded Average Level Scores		
	Number of Examinees	$\hat{\sigma}_l(\Delta)$	$\hat{\sigma}_l(\delta)$	Number of Examinees	$\hat{\sigma}_l(\Delta)$	$\hat{\sigma}_l(\delta)$
0	121	0.52011	0.50862	19	0.34847	0.33357
1	300	0.93071	0.92434	132	0.51769	0.50778
2	1785	0.55617	0.54544	1167	0.39277	0.37962
3	3682	0.35216	0.33496	3259	0.35331	0.33863
4	1203	0.34791	0.33048	2471	0.33327	0.31766
5	6	0.40472	0.38985	49	0.42024	0.40797

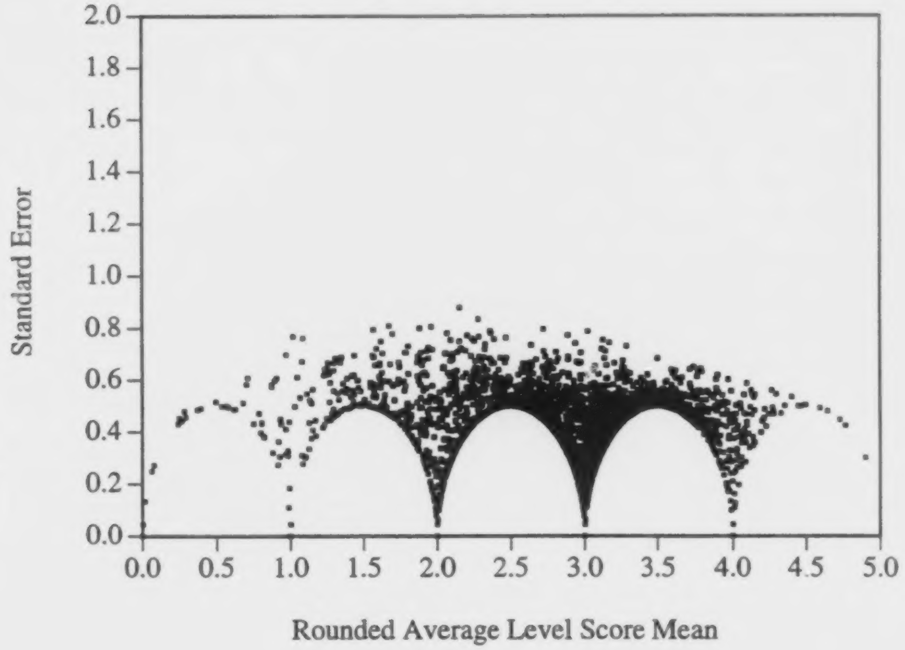

Figure 1. Absolute conditional standard errors for Listening Rounded Average Level Scores.

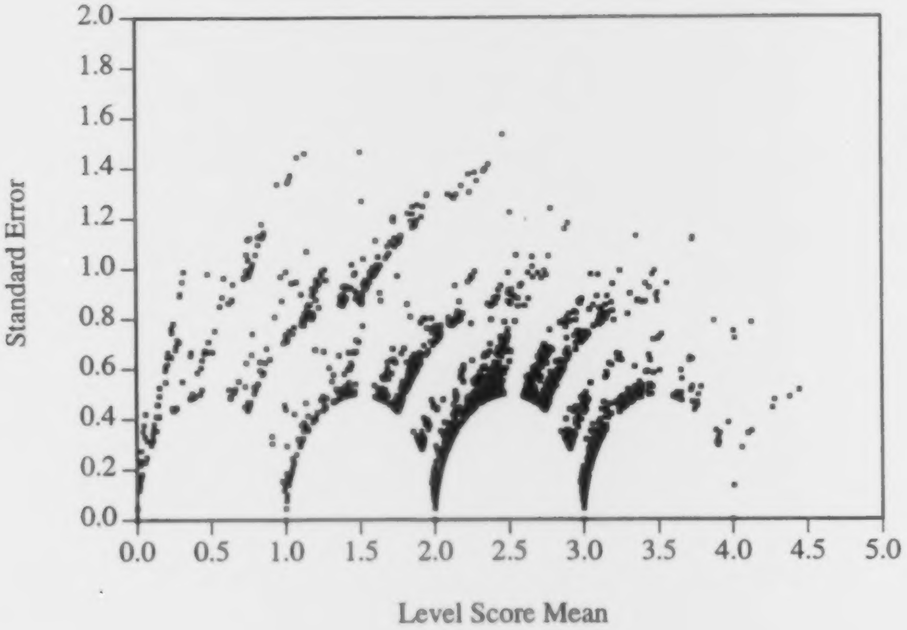

Figure 2. Absolute conditional standard errors for Listening Level Scores.

BEST COPY AVAILABLE

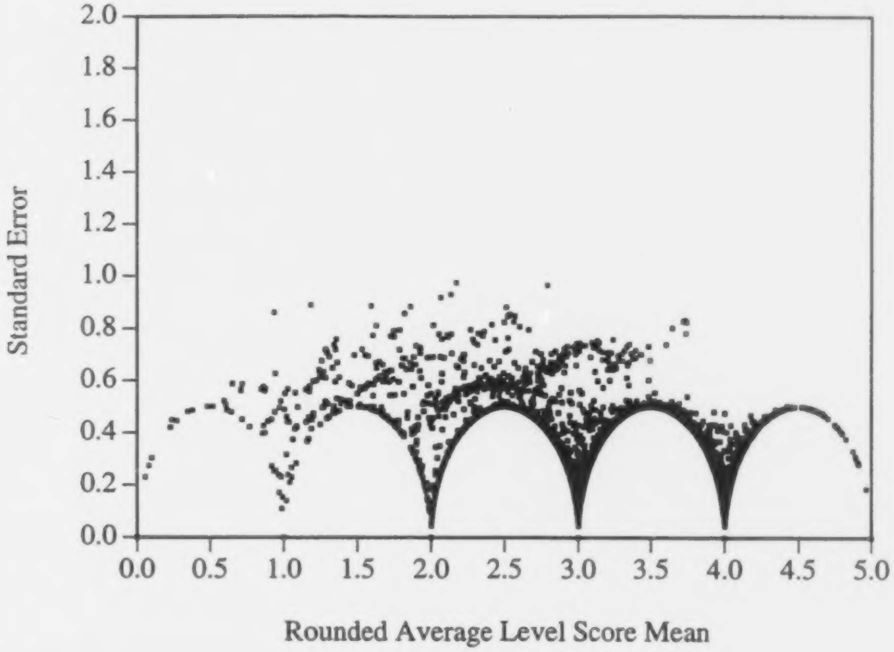

Figure 3. Absolute conditional standard errors for Writing Rounded Average Level Scores.

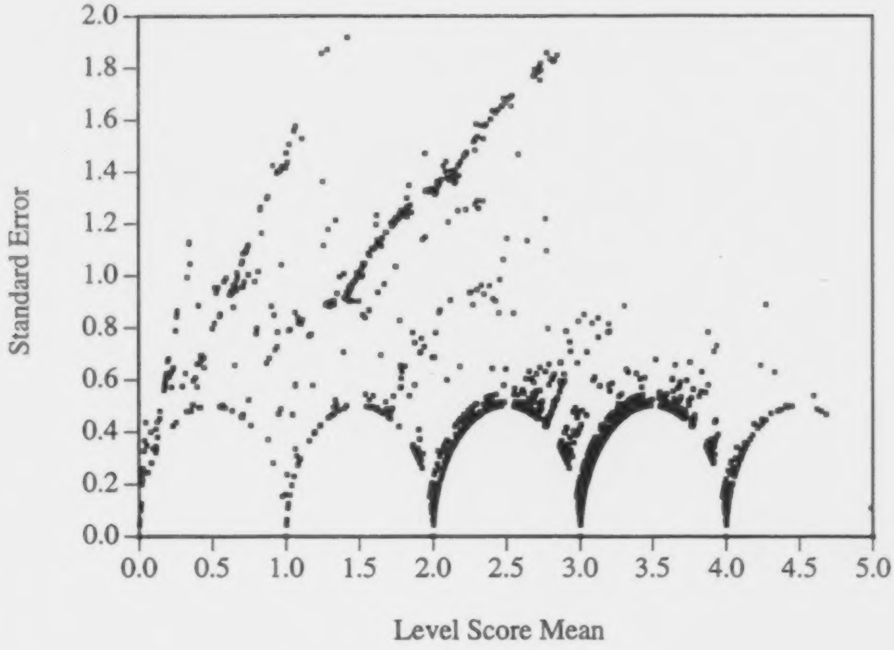

Figure 4. Absolute conditional standard errors for Writing Level Scores.

Evaluating Measurement Precision of Performance Assessment With Multiple Forms, Raters, and Tasks

Xiaohong Gao and Dean A. Colton

Abstract

Single-form scores are likely to be used to judge individuals' performance levels due to high cost in performance assessments. However, it is not clear whether estimates of individual performance are consistent from one test form to another. If people are willing to make decisions based on a single-form score, it is important to know the score generalizability across forms. The purpose of the present study was to examine measurement precision of performance scores when multiple forms, raters, and tasks were used in the measurement.

Moreover, raw scores are usually non-linearly transformed into scale scores. Little research has been done about measurement precision of such scores. A bootstrapping method combined with generalizability analyses was used to estimate conditional standard errors of measurement and generalizability of scale (level) scores. The results indicate that (a) examinees' scores vary from one form to another; (b) within a form, the rank ordering of task difficulty is substantially different for the various examinees; (c) measurement errors are mainly introduced by task sampling variability not by rater sampling variability; (d) writing scores are more generalizable than listening scores; and (e) level (scale) scores are less generalizable than average raw scores.

Evaluating Measurement Precision of Performance Assessment With Multiple Forms, Raters, and Tasks*

Research on the sampling variability and generalizability of performance assessments has indicated that (a) an individual's performance score varies greatly from one task to another, (b) a large number of tasks are needed to obtain a generalizable measure of an individual's performance, and (c) well-trained raters can provide reliable ratings (Brennan, Gao, & Colton, 1995; Gao, Shavelson, & Baxter, 1994; Shavelson, Baxter, & Gao, 1993). However, in most performance assessments, an individual takes only one test form due to resource constraints, and a single form score is likely to be used to make judgments about the individual's performance. With a narrower universe than the one to which generalization is likely to be made, measurement errors are likely to be underestimated.

A test form is a collection of test items (tasks) and is built according to certain content and statistical specifications. Although test developers attempt to assemble test forms as parallel (equivalent) as possible they usually differ somewhat in difficulty and contribute to sampling variability. In some performance assessments, equating may not be conducted to adjust for differences in difficulty among forms. It is also not clear whether an individual's performance scores are consistent from one test form to another. If there is a large person by form interaction, conventional equating methods may not be applicable. Under these circumstances, can test forms designed to measure the same construct be used interchangeably? If people are willing to make decisions or judgments about individuals based on single-form scores without any score adjustment, it is essential to investigate sampling variability across forms. Furthermore, when multiple raters and tasks are used, in addition to multiple forms, it is important to examine the magnitude of sampling variability associated with those sources and their impact on measurement errors and generalizability.

---

* The authors gratefully acknowledge the contributions of Robert L. Brennan to the design of the original study and his comments on an earlier version of the paper. We also express our appreciation for the comments and suggestions of Michael J. Kolen and Deborah J. Harris.

In practice, raw scores of a test are usually non-linearly transformed into scale scores (e.g., proficiency or level scores) which are reported to examinees. Naturally, it is essential to estimate measurement errors, especially conditional standard errors of measurement, and reliability associated with scale scores. Although extensive research has been done about measurement precision of raw scores, literature on issues related to scale-scores is scarce (but see Brennan & Lee, 1997; Colton, Gao, & Kolen, 1996; Feldt & Qualls, in press).

The purpose of the present study was to examine sampling variability and generalizability of a performance-based listening and writing assessment with multiple forms, raters, and tasks involved. More specifically, the study addresses the following questions: (a) What are the major sources of measurement errors associated with the measurement procedure used in the assessment: forms, raters, and/or tasks? (b) What are the effects of changing measurement procedures (e.g., using different numbers of raters, tasks, and/or forms) on measurement errors? (c) What are the effects of changing measurement procedures on score generalizability? (d) What are the conditional standard errors of measurement (CSEM) at different performance levels? and (e) How generalizable are aggregated proficiency (level) scores? Generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 1992; Shavelson & Webb, 1991) and a bootstrap method (Efron & Tibshirani, 1993) can be brought to bear on these technical issues.

Method

Data

The data were collected in a 1993 study conducted by ACT, Inc. to evaluate the Work Keys assessment system. Two forms (A and B) of the Work Keys Listening and Writing assessment were administered to 167 examinees. Although a plan was made to counter-balance the two test forms, the procedure was not followed strictly during the test administration. For a given form, three raters assigned Listening scores to all six tasks (prompts) for all examinees. A different group of three raters assigned Writing scores to all six tasks for the examinees. The groups of raters were also different for each form.

Instrument

An important feature of the Work Keys Listening and Writing assessment is that a single set of tasks (prompts) are administered, but two different performance scores--Listening and Writing--are provided. The Listening score indicates an examinee's skill at listening to and understanding work-related messages, whereas the Writing score indicates the examinee's skill at composing and writing work related messages. The assessment was administered via an audio tape that contained all directions and messages (prompts). Examinees were asked to listen to six audio-taped messages ranging from shorter and easier to longer and more complex. After listening to each recorded message, examinees were told to construct a written summary of the prompt. The written responses were scored once for listening and again for writing skills. The Listening score was based on the accuracy and completeness of the information in the examinee's written responses, and the Writing score was based on the writing mechanics (such as sentence structure and grammar) and writing style used in the examinee's written responses. All scoring was done by three raters in the situation reported here. In the usual operational scoring, only two raters are used. The raw scores ranged from 0 to 5 for each task.

Design and Analysis

The analyses were carried out in two parts: both sample estimates and bootstrap estimates of sampling variability and generalizability were generated. In the sample-estimation part, two designs were used to conduct generalizability analyses of the Work Keys Listening and Writing assessment: person x [(rater x task):form] and person x form. In the bootstrap-estimation part, a person x form design was used. The analyses examined sampling variability, standard error of measurement and score generalizability. Separate analyses were conducted for Listening and Writing.

Sample estimation. A performance assessment score is subject to sampling variability. An important contribution of generalizability theory to measurement is that it allows researchers to disentangle multiple sources of measurement error associated with various measurement procedures. The purpose of the Work Keys study was to examine sources of variability related to forms, raters, and tasks. The original data collection design contained three facets--forms, raters, and tasks. More specifically, the examinees (p) took two test forms (f), each form contained six tasks (t), and each written response was scored by three raters (r). The linear model and variance components for the design are presented under the first four headings in Table 1. (The bootstrap formulas shown under the last two headings in Table 1 are described later.) The generalizability analysis allows us to examine the magnitudes of sampling variabilities of forms, raters, and tasks, as well as the interactions.

Work Keys reports aggregated proficiency scores (Level Scores) for Listening and Writing to individuals and to educational and business agencies. Operationally, two raters score each of the six tasks. An examinee receives a particular level score if at least 9 of the 12 ratings (6 tasks x 2 raters) are at or higher than that score. Since there were three raters in the present study which generated three pairs of raters (i.e., raters 1 and 2, raters 1 and 3, raters 2 and 3), three level scores were assigned to each person. To use all these ratings in the analyses these three Level scores were then averaged and rounded to represent each person's level score (0-5). In addition, rounded mean raw scores averaged over the three raters and six tasks were also calculated for the examinees on each form (i.e., Rounded-Average Level Scores). Person x form generalizability analyses were carried out using Level Scores and Rounded-Average Level Scores to examine form-sampling variability when aggregated scores were used (see Table 1). The measurement precision of the two types of aggregated scores were also compared.

Decision (D) studies, more precisely D-study considerations, with various numbers of conditions were then conducted using the two designs: p x [(R x T):F] and p x F. The uppercase letters are indices of the sources of variability in the D-study considerations which are used interchangeably in this report. Since the Work Keys assessment scores are used to index individual performance levels, the present study focuses on absolute error variances and absolute generalizability (G) coefficients or dependability coefficients (Φ) (see Table 1).

TABLE 1

Equations for Generalizability Analyses

Linear Models

$$X_{(prt:f)} = \mu + \mu_{p\sim} + \mu_{f\sim} + \mu_{r:f\sim} + \mu_{t:f\sim} + \mu_{pf\sim} + \mu_{pr:f\sim} + \mu_{pt:f\sim} + \mu_{rt:f\sim} + \mu_{prt:f\sim}$$

$$X_{pf} = \mu + \mu_{p\sim} + \mu_{f\sim} + \mu_{pf\sim}$$

G-Study Total Variances

$$\sigma^2(X_{prt:f}) = \sigma^2_p + \sigma^2_f + \sigma^2_{r:f} + \sigma^2_{t:f} + \sigma^2_{pf} + \sigma^2_{pr:f} + \sigma^2_{pt:f} + \sigma^2_{rt:f} + \sigma^2_{prt:f}$$

$$\sigma^2(X_{pf}) = \sigma^2_p + \sigma^2_f + \sigma^2_{pf}$$

Absolute Error Variances

$$\sigma^2_\Delta = \frac{\sigma^2_f}{n'_f} + \frac{\sigma^2_{r:f}}{n'_{r:f}n'_f} + \frac{\sigma^2_{t:f}}{n'_{t:f}n'_f} + \frac{\sigma^2_{pf}}{n'_f} + \frac{\sigma^2_{pr:f}}{n'_{r:f}n'_f} + \frac{\sigma^2_{pt:f}}{n'_{t:f}n'_f} + \frac{\sigma^2_{rt:f}}{n'_{r:f}n'_{t:f}n'_f} + \frac{\sigma^2_{prt:f}}{n'_{r:f}n'_{t:f}n'_f}$$

$$\sigma^2_\Delta = \frac{\sigma^2_f}{n'_f} + \frac{\sigma^2_{pf}}{n'_f}$$

Dependability Coefficient

$$\Phi = \frac{\sigma^2_{\mathrm{p}}}{\sigma^2_{\mathrm{p}} + \sigma^2_\Delta}$$

Standard Error of Measurement (Individual)

$$\sigma(\Delta_{\mathrm{p}}) = \sigma(\mathrm{X}_{\mathrm{pb}}) = \sqrt{\frac{\sum_{\mathrm{b}=1}^{\mathrm{n_b}} \mathrm{X}^2_{\mathrm{pb}} - (\sum_{\mathrm{b}=1}^{\mathrm{n_b}} \mathrm{X}_{\mathrm{pb}})^2 / \mathrm{n_b}}{\mathrm{n_b} - 1}}$$

Conditional Standard Error of Measurement (Level)

$$\sigma_{\mathrm{l}}(\Delta) = \sqrt{\frac{1}{\mathrm{n_{p:l}}} \sum_{\mathrm{p}=1}^{\mathrm{n_{p:l}}} \sigma^2(\Delta_{\mathrm{p}})}$$

Note. The equations for calculating individual standard error of measurement and conditional standard error of measurement in generalizability theory were initially discussed by Brennan (1996).

BEST COPY AVAILABLE

Bootstrap estimation. The p x f generalizability analysis in the previous section provided only sample estimates of variance components, error variances, and generalizability coefficients for level scores. A bootstrap method (Efron & Tibshirani, 1993) can be used to further estimate an individual's standard error of measurement and conditional standard error of measurement (CSEM) at each performance level. A generalizability analysis using the p x f bootstrapping data provides estimates of variance components and generalizability coefficients for the level scores.

In a bootstrap analysis, repeated samples (usually 500 or more) are drawn from a data matrix. This data matrix is treated "as if" it were the population and repeated samples with replacement are drawn from it. Statistics (e.g., means) are calculated from each bootstrapping sample and the stability of parameter estimates (i.e., standard deviations) can be computed. In the present study, two raw data matrices, one for Listening and another for Writing, were used as data bases for the simulation and bootstrap analyses. Each matrix contained 167 persons by 24 scores (6 tasks, 2 raters, and 2 forms). More specifically, the 167 examinees took two forms and their 12 written responses were scored by two raters. If the two raters disagreed by more than one point a third rater's score was used to replace the original two raters' scores. (This procedure is referred to resolution by an expert rater in operational administrations). Raters for the two forms were different.

The bootstrap procedure was carried out separately for Listening and Writing in the following steps:

1. A form was randomly selected and the first task in the form was used. A form was randomly chosen again and the second task from that form was selected. This process was repeated six times until the sixth task was sampled to create a bootstrap form that would have a similar structure as the original forms (i.e., six tasks were ordered from easy to difficult and from short to long).
2. A 167 x 12 bootstrap sample was created using the examinees' scores on these selected six tasks given by two raters.

3. Steps 1 and 2 were repeated 500 times to create 500 bootstrap samples (replications), each containing a 167 x 12 matrix.
4. Bootstrap Level Scores and Rounded-Average Level Scores were computed for the examinees in each of the bootstrap samples. These scores were based on two raters' ratings only. Consequently, 500 Level Scores and 500 Rounded-Average Level Scores were computed for each examinee.

The two new 167 x 500 data sets, one containing Level Scores and another containing Rounded-Average Level Scores, were used to calculate standard deviations of the bootstrap estimates. These standard deviations were considered as estimates of the standard error of measurement (or absolute error) for each examinee. The examinees were then assigned to six groups according to their average scores over the 500 bootstrap replications. The average standard errors at each of the six levels were computed and considered as estimates of absolute CSEM.

In addition, a person x form (i.e., bootstrap sampled forms) random-effects generalizability analysis was carried out to estimate variance components for person, form, and person by form interaction, measurement errors and dependability coefficients for Listening and Writing. Equations for these conditional standard errors are provided under the last two headings in Table 1. In these equations, the "b" subscript refers to a bootstrap replication, n_b is the number of bootstrap replications, $n_{p:l}$ is the number of persons nested within levels.

Results and Discussion

Sampling variability and generalizability of the Work Keys Listening and Writing assessment were examined using generalizability theory and the bootstrap method. A series of generalizability (G) analyses were conducted to (a) estimate variance components associated with various sources of sampling variation (i.e., form, rater, and task), (b) assess standard errors of measurement for different measurement procedures (i.e., different numbers of conditions in the facets), and (c) examine the generalizability of the Listening and Writing assessment scores. Besides the sample estimates, the bootstrap method was used in conjunction with the use of generalizability theory to estimate conditional standard errors of measurement and generalizability coefficients. The results provide information about likely psychometric characteristics of the assessment.

Sample Estimates of Measurement Precision

Estimated variance components . Table 2 provides the estimated G-study variance components, $\hat{\sigma}^2(\alpha)$, for the p x [(r x t):f] design and the associated percents of total variability (%) for Work Keys Listening and Writing scores. The estimates indicate the magnitudes of sampling variation associated with each source (forms, raters, and tasks) and their relative contributions to measurement errors. The person by task interaction contributes most to measurement errors for both Listening and Writing, indicating that the rank orders of examinees vary from one task to another. The finding of a large person by task interaction is consistent with other reported results on performance assessments (see Brennan et al., 1995; Gao et al., 1994; Shavelson et al., 1993). Moreover, the estimated task variance component is the second largest for Listening, suggesting that the tasks within a form differ in difficulty. The task means for Listening Form A range from 2.383 to 3.473. The results are consistent with the test descriptions which state that the tasks are ordered from easy to difficult. However, tasks do not.differ so greatly in difficulty for Writing. The means for Writing Form A range from 2.764 to 3.200. The most notable difference in the results for Listening and Writing is the (t:f) component: Listening score is affected by task complexity, but one's ability to construct a good sentence is not.

Further, the form difficulty, averaging over examinees, raters, and tasks, is different for Listening but not for Writing. For example, the mean is 2.795 for Listening Form A but is 3.226 for Listening Form B. The average Writing scores are 2.977 for Form A and 2.913 for Form B, respectively. However, the individual scores vary somewhat from one form to another (i.e., person by form interaction) for both Listening and Writing. The results suggest that some score adjustment may be needed so that the Listening and Writing scores obtained from different forms are comparable.

TABLE 2

Variance Component Estimates of the p x [(r x t):f] Design

	Listening		Writing	
Source of Variability	$\hat{\sigma}^2(\alpha)$	%	$\hat{\sigma}^2(\alpha)$	%
Person (p)	0.26104	21.04	0.37201	45.83
Form (f)	0.04529	3.65	0.00000	0.00
Rater:form (r:f)	0.00472	0.38	0.00410	0.51
Task:form (t:f)	0.26973	21.75	0.01136	1.40
pf	0.01767	1.42	0.01964	2.42
pr:f	0.00755	0.61	0.01908	2.35
pt:f	0.47268	38.11	0.23229	28.61
rt:f	0.00338	0.27	0.00353	0.43
prt:f	0.15833	12.76	0.14976	18.45

For Writing, the universe score (true score) variance is larger than the other estimated variance components and is larger than that for Listening, suggesting that there is considerably more variation among examinees with respect to their levels of proficiency in writing than in listening. Similar findings were reported on Work Keys data collected in a previous year (see Brennan et al., 1995).

As seen in Table 2, the rater-sampling variability is small, especially for Listening. The fact that rater variance is small means that raters are about equally stringent on average. The fact that the rater-by-person interaction is small means that examinees are rank ordered about the same by the various raters. The results, thus, suggest that raters are not nearly as large a contributor to total variance as are tasks. It is possible to use a small number of well-trained raters to score each examinee's responses in future operational forms if the training and scoring procedures continue to be well developed and used. It is noteworthy that the variance component (prt:f) for a person by rater by task interaction confounded with other unidentified sources of error is relatively large.

The estimates in Table 2 are for single person-rater-task-form scores only. In practice, decisions about examinees are typically made based on average or total scores over some numbers (n') of tasks, raters and/or forms defined by a universe of generalization. Assuming one form, two raters and six tasks are used in the p x [(R x T):F] D-study considerations, Table 3 provides the estimated variance components for the Listening and Writing assessment. Increasing the number of tasks from one to six dramatically decreases the estimated task variance components, and the person by task interactions for both Listening and Writing although tasks still count for a large proportion of the total variability.

TABLE 3

Variance Component Estimates of the p x [(R x T):F] Design

	Listening		Writing	
Source of Variability	$\hat{\sigma}^2(\alpha)$	%	$\hat{\sigma}^2(\alpha)$	%
Person (p)	0.26104	55.86	0.37201	81.47
Form (F)	0.04529	9.69	0.00000	0.00
Rater:form (R:F)	0.00236	0.50	0.00205	0.45
Task:form (T:F)	0.04496	9.62	0.00189	0.41
pF	0.01767	3.78	0.01964	4.30
pR:F	0.00378	0.81	0.00954	2.09
pT:F	0.07878	16.86	0.03871	8.48
RT:F	0.00028	0.06	0.00029	0.06
pRT:F	0.01319	2.82	0.01248	2.73

The above generalizability analysis was conducted on raw scores of the Listening and Writing assessment. The p x f generalizability analysis dealt with Level Scores transformed non-linearly from raw scores and Rounded-Average Level Scores. As indicated in the top part of Table 4, the form variability is notably larger for Listening than for Writing, indicating that the two forms are not equivalent in average difficulty for the Listening test. The form variance component estimates for Writing are negligible. The results are consistent with those reported earlier in the p x [(r x t):f] generalizability analysis with raw scores. Moreover, the large person by form interactions for both Listening and Writing scores suggest that the rank orders of examinees vary by forms.

Estimated standard errors of measurement. For the measurement procedure used in the original data collection (i.e., $n_r = 3$, $n_t = 6$, and $n_f = 2$) the measurement errors are smaller for Writing (0.20) than for Listening (0.32). Figure 1 at the end of this report demonstrates that standard errors of measurement, $\hat{\sigma}(\Delta)$, are reduced when D-study sample sizes (n'_r, n'_t, and n'_f) increase. Although increasing the number of raters doesn't improve the measurement precision very much, especially for Listening, adding more tasks and/or forms does. In the p x F D-study with $n'_f = 1$, the standard errors, $\hat{\sigma}(\delta)$ for relative decisions and $\hat{\sigma}(\Delta)$ for absolute decisions, are smaller for Writing than for Listening and are smaller for the Rounded-Average Level Scores than for the Level Scores (see the sample estimates in Table 4).

In practice, the estimated standard errors of measurement, $\hat{\sigma}(\Delta)$, can be used to construct the confidence intervals (or bands) that are likely to contain universe (true) scores, assuming that errors are normally distributed (Cronbach, Linn, Brennan, Haertel, 1997). The 90% confidence interval containing an examinee's true performance level would be in the range of $\pm 1.645\,\hat{\sigma}(\Delta)$. For example, with a $\hat{\sigma}(\Delta) = 0.62$, the interval for the Listening Level Scores is ± 1.02, and with a $\hat{\sigma}(\Delta) = 0.51$ the interval for the Rounded-Average Level Scores is ± 0.84. Likewise, the Listening scores have wider confidence intervals than the Writing scores due to the larger standard errors. In addition, $\hat{\sigma}(\Delta)$ can provide information on the probability (or the percentage) of misclassification of the examinee(s) and can be used to estimate minimum passing and maximum failing scores given a specified standard of proficiency with a certain level of confidence (Linn & Burton, 1994).

TABLE 4

Variance Component Estimates of the p x f Design

	Listening		Writing	
Source	Level	Rounded	Level	Rounded
Sample Estimates				
Person (p)	0.27151	0.26312	0.44643	0.40608
Form (f)	0.11221	0.08754	0.00052	0.00008
pf	0.27202	0.17394	0.20745	0.16590
$\hat{\sigma}(\delta)$	0.52156	0.41706	0.45547	0.40731
$\hat{\sigma}(\Delta)$	0.61986	0.51135	0.45604	0.40741
$E\hat{\rho}^2$	.50	.60	.68	.71
$\hat{\Phi}$	.41	.50	.68	.71
Bootstrap Estimates				
Person (p)	0.33180	0.30124	0.48500	0.45026
Form (f)[a]	0.01865	0.02794	0.00061	0.00118
pf	0.25987	0.17934	0.15648	0.12519
$\hat{\sigma}(\delta)$	0.50977	0.42349	0.39558	0.35382
$\hat{\sigma}(\Delta)$	0.52775	0.45528	0.39633	0.35550
$E\hat{\rho}^2$	.56	.63	.76	.78
$\hat{\Phi}$	.54	.59	.76	.78

[a] Bootstrap replications.

Estimated generalizability coefficients . The generalizability (G) coefficients for the p x [(R x T):F] design depend, in part, upon the numbers of raters (n'_r), tasks (n'_t), and/or forms (n'_f) used in decision considerations. If only one form, two raters, and six tasks were used (see Table 3 for the variance component estimates), the absolute G coefficient or dependability coefficient ($\hat{\Phi}$) would be .56 for Listening and .81 for Writing. Figure 1 demonstrates that dependability coefficients (PHI) increase when D-study sample sizes (n'_f, n'_r, and n'_t) increase. However, increasing the number of raters beyond two doesn't improve the score generalizability substantially, especially for Listening; but adding more tasks and/or forms does.

In the p x F D-study, the generalizability coefficient ($E\hat{\rho}^2$) and dependability coefficient ($\hat{\Phi}$) for Writing are higher than those for Listening (see the sample estimates in Table 3), suggesting that the Writing scores are more generalizable than the Listening scores for relative and absolute decisions. In addition, the Rounded-Average Level Scores are more generalizable than the Level Scores.

Bootstrap Estimates of Measurement Precision

Generalizability estimates. The bottom of Table 3 presents the bootstrap variance component estimates, standard errors of measurement, $\hat{\sigma}(\delta)$ (relative error) and $\hat{\sigma}(\Delta)$ (absolute error), generalizability ($E\hat{\rho}^2$) and dependability coefficients ($\hat{\Phi}$) for both Level Scores and Rounded-Average Level Scores. The results have similar patterns as those from sample estimates: the Writing test is more generalizable than the Listening test; the Rounded-Average Level Scores are more generalizable than the Level Scores. They are consistent with findings from a study conducted by Colton, Gao, and Kolen (1996) using a different Work Keys data set.

It is noteworthy that the bootstrap estimates of the universe-score variance are larger than the estimates based upon the generalizability analysis of the sample. The differences in the magnitudes of these estimates may be partly due to the bootstrap sampling procedure used in the study. Brennan, Harris, and Hanson (1987) show that the variance component for persons is likely to be overestimated in a person x item design when only items are bootstrapped.

Conditional standard errors of measurement. Table 5 reports the estimated CSEM of bootstrap Level Scores and Rounded-Average Level Scores. The estimated standard errors for the Rounded-Average Level scores tend to be lower than those for the Level Scores in both Listening and Writing (see also Colton et al., 1996). The Writing CSEMs are lower than Listening CSEMs at Levels 2, 3, and 4. The CSEM estimates are not stable at the extreme score levels due to small sample sizes.

TABLE 5

Conditional Standard Errors of Measurement

	Listening				Writing			
	Level Scores		Rounded Average		Level Scores		Rounded Average	
Level	n	CSEM	n	CSEM	n	CSEM	n	CSEM
0	2	0.50412	0	N/A	2	0.93333	0	N/A
1	7	0.67571	1	0.54304	4	0.85620	0	N/A
2	64	0.51874	17	0.49102	62	0.38652	44	0.36508
3	91	0.51243	112	0.44758	80	0.34227	78	0.35771
4	3	0.75096	36	0.45771	19	0.39230	45	0.34188
5	0	N/A	1	0.49372	0	N/A	0	N/A

Conclusions

The generalizability and bootstrap analyses reported here reveal that (a) examinees' scores vary from one test form to another which may be partly due to large task-sampling variability, (b) the rank orderings of task difficulty differ across the examinees, (c) measurement errors are mainly introduced by task-sampling variability not by rater-sampling variability, (d) the Writing scores are more generalizable than the Listening scores, and (e) Level Scores are less generalizable than Rounded-Average Level Scores. The results portray some important psychometric properties about Work Keys Listening and Writing scores.

The finding that examinees are rank ordered differently on different forms of the Listening test suggests that measurement errors are likely to be underestimated in situations where individuals take only one test form. Further, score adjustments may be needed to make scores generated from different forms comparable in making decisions. However, conventional equating methods may not be entirely satisfactory here due to the person-by-form interaction. The result that examinees' performances vary from one task to another is consistent with other findings in performance assessments. These findings strongly indicate the importance of domain specification and task sampling in test development (Shavelson, Gao, & Baxter, 1995). The finding that one or two well-trained raters can reliably score examinees' performance is encouraging for future test operations.

The present study combines generalizability theory and the bootstrap method to examine sampling variability, conditional standard errors of measurement, and generalizability of scale (level) scores. These methods may be used to evaluate technical qualities in other performance-assessment situations where a single score is used to index individuals' levels of performance (absolute decisions) or to rank order individuals (relative decisions). The bootstrap method can be used to generate level scores for examinees using their individual raw scores (see Colton et al., 1996). Generalizability analyses can then be used to estimated conditional standard errors of measurement and generalizability coefficients for both relative and absolute decisions.

References

Brennan, R. L. (1992). *Elements of generalizability theory* (rev. ed.). Iowa City, IA: American College Testing.

Brennan, R. L. (1996). *Conditional standard errors of measurement in generalizability theory.* (ITP Occasional Paper No. 40). Iowa City, IA: Iowa Testing Programs, University of Iowa.

Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analyses of Work Keys Listening and Writing Tests. *Educational and Psychological Measurement, 55* (2), 157-176.

Brennan, R. L., Harris, D. J., & Hanson, B. A. (1987, April). *The bootstrap and other procedures for examining the variability of estimated variance components in testing contexts.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, D. C.

Brennan, R. L., & Lee, W. C. (1997). *Conditional standard errors of measurement for scale scores using binomial and compound binomial assumptions.* (ITP Occasional Paper No. 41). Iowa City, IA: Iowa Testing Programs, University of Iowa.

Colton, D. A., Gao, X., & Kolen, M. J. (1996). Assessing the reliability of performance level scores using bootstrapping. In M. J. Kolen (Chair), *Technical issues involving reliability and performance assessments.* Symposium conducted at the Annual Meeting of the American Educational Research Association, New York.

Cronbach, L. J., Gleser, G. C., Nanda, H. I., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York: Wiley.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*(3), 373-399.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap* (Monographs on Statistics and Applied Probability 57). New York: Chapman & Hall.

Feldt, L. S., & Qualls, A. L. (in press). Approximating scale score standard error of measurement from the raw score standard error. *Applied Measurement in Education.*

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education, 7* (4), 323-342.

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice, 13* (1), pp. 5-8, 15.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30* (3), 215-232.

Shavelson, R. J., Gao, X., Baxter, G. P. (1995). On the content validity of performance assessments: Centrality of domain specification. In M. Birenbaum & F. Dochy (Eds.), *Alternatives in assessment of achievements, learning process and prior knowledge* (pp. 131-141). Boston: Kluwer Academic Publishers.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage Publications.

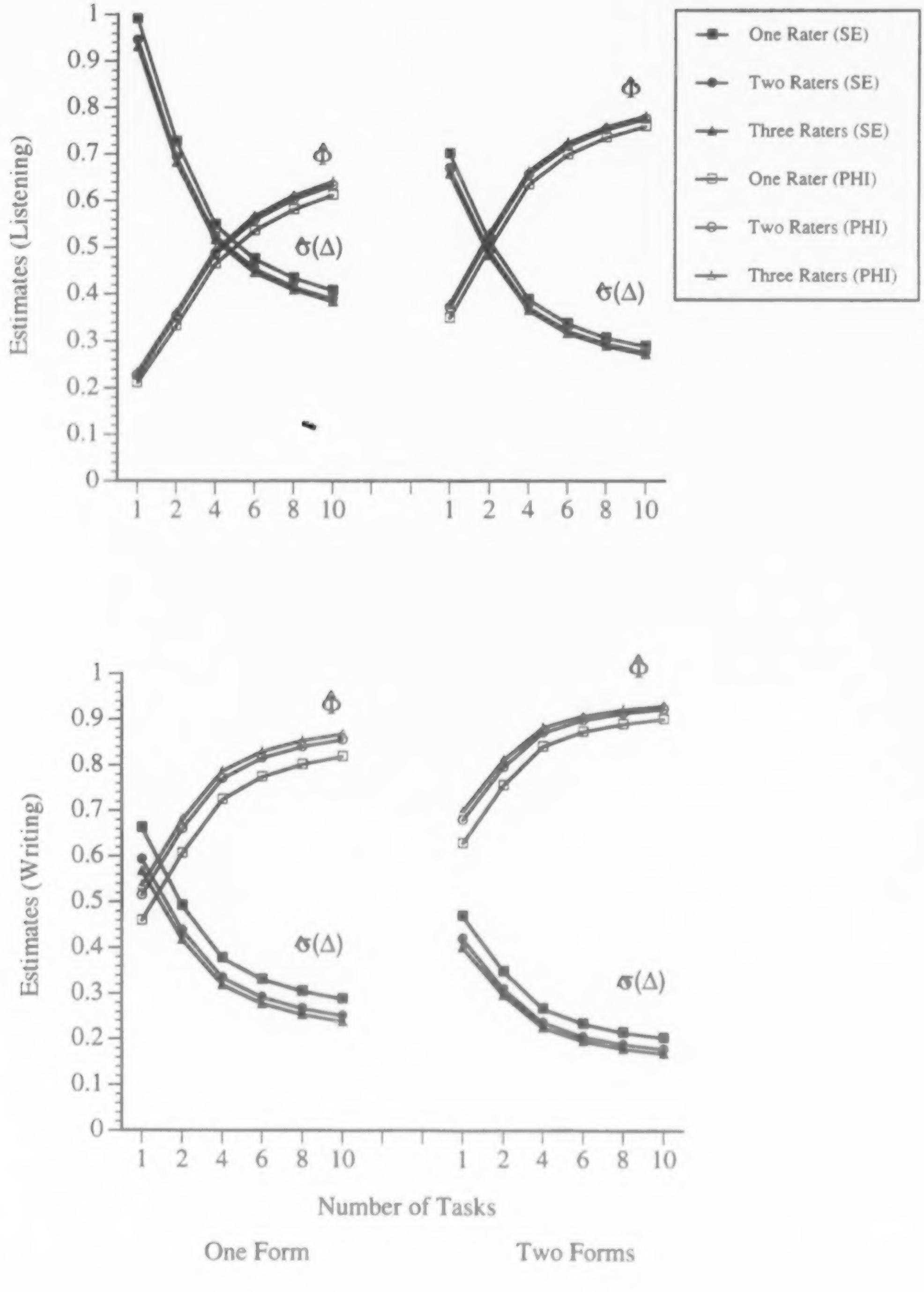

FIGURE 1. Estimated absolute errors and dependability coefficients for p x [(R x T):F] D-study considerations

Weights that Maximize Reliability Under a Congeneric Model for Performance Assessment

Tianyou Wang

Abstract

In many measurement situations, there is often a need to weight different scores to form a composite score. One important factor in determining the weights may be the desirability to maximize the reliability of the composite score. Procedures derived in the past usually require information about the reliability of each part score and are computationally complex. In some situations, reliability estimates of the part scores may not be readily available. In this paper, formulas for computing the weights that maximize the reliability of a test with multiple parts are derived using a congeneric model. A direct derivation for the three-part case and a two-step derivation for the *n*-part case are presented and results for these two approaches are shown to be consistent for the three-part case. The formulas are rather simple and are all based on the variance-covariance matrix of the part scores. Two examples are given to show the computations and the usefulness of the formulas.

Weights that Maximize Reliability Under a Congeneric Model for Performance Assessment

In educational measurement, it is not uncommon to assign weights to several scores when it is needed to combine them to form a composite score (Wang & Stanley, 1970; Feldt & Brennan, 1989, p. 116). Combining scores to form a composite score may occur at different levels. At the highest level in one possible direction of ordering, we may combine scores from different tests built on different score scales. This happens when we use different test scores and school grades to predict, say, college grade point average (GPA) using a multiple regression model. The predictor is in effect a weighted composite of these individual predictors with the weights being the regression coefficients . At lower levels, we may combine scores from tests within a test battery, or parts scores within a test. With this ordering, the lowest level may be the individual test item level.

In conventional multiple-choice based testing, although the typical practice is to use unweighted number-correct raw scores as the basis for scoring, it is possible to assign empirically derived weights to form some kind of optimal scores. For instance, Lord (1980, pp. 73-77) discussed ways of using item response theory (IRT) based item statistics as weights for optimal scoring. An emerging form of testing situation that may make combining scores at the item level a useful practice is performance assessment. Performance assessment instruments typically consist of more than one task or prompt. The tasks are typically of unequal depth or difficulty level. So it is a common practice to assign some weights to those task scores to form a composite score. The weighting scheme can be designed based on different considerations such as content validity (importance), testing time, and reliability. Though the final decision about the weights may depend on several different factors, empirically derived weights based on certain criteria can be used to facilitating the process of choosing weights. One such criterion can be the reliability of the composite score; i.e., to find a set of weights to maximize the reliability of the composite score. (For simplicity in expression, the weights that maximize reliability will be referred to as the optimal weights.) The purpose of this paper is to derive formulas for such weights that are potentially useful in performance assessments and other assessment settings.

Previous research has studied ways of finding the optimal weights of the composite score for tests that have subtests. This research can be organized into three categories. The first category (Thomson, 1940; Mosier, 1943; Peel, 1947) uses the classical test theory formulation for composite score reliability for a test battery. In their approaches, the reliability of each subtest needs to be estimated using classical test theory methods such as internal consistency indices. Li, Rosenthal, and Rubin (1996) and Li (1997) presented expressions for the maximum reliability for this category of methods. The second category (Joe & Woodward, 1976; Conger, 1974) uses multivariate generalizability theory to find weights that maximize the reliability of the test battery composite score. Solutions for these two categories involve finding the largest eigenvalue and the associated eigenvector for a certain matrix, and so the solution does not have a closed form. Also, these methods generally require that the subtests contain multiple items in order to estimate subtest reliability or variance components. The third category (Kaiser & Caffrey, 1965; Bentler, 1968) uses factor analytical methods to find weights for each individual item to maximize the internal consistency reliability of the test. These methods involve estimating communalities and do not have a closed form because they also require solving for the largest eigenvalue and associated eigenvector for a certain matrix.

The present paper aims to address a situation where there is no information about the reliability of each part score. The only available data may be the variance-covariance matrix of the part scores. This situation makes the first two categories of the previously described methods difficult to implement. The factor analytical methods may also not be very practical because of the computational complexity. Practitioners often need some simple straight forward formulas that can be used to compute reasonably good estimates of the optimal weights to help them to decide the final weights.

The procedure proposed here assumes a congeneric model for the test. The congeneric model states that the items are measuring the same construct but that their true score variances and error score variances may differ. Reliability here is defined as the correlation of the scores derived from parallel forms. The parallel forms are supposed to use the same score weighting.

Define X as the total observed score on a test that has n parts, and X_i and E_i as, respectively, the observed score and error score for part i. Under the congeneric model, the test scores can be expressed as follows:

$$
\begin{aligned}
&X = X_1 + X_2 + \cdots + X_n \text{, where} \\
&X_1 = \lambda_1 T + E_1 \\
&X_2 = \lambda_2 T + E_2 \\
&\vdots \\
&X_n = \lambda_n T + E_n,
\end{aligned}
$$

and the λ_i's are the congeneric coefficients.

It is assumed that all of the λ_i's are positive (if not, one can reverse the scale to make them positive), that $\sum_i \lambda_i = 1$, and that all of the error variances, $\sigma^2_{E_i}$'s, do not have to be equal. The reliability formulas under a congeneric model are summarized in Feldt and Brennan (1989, p. 115). The expression for the reliability coefficient is much more complicated when the test has more than three parts than when it has only three parts. Therefore alternative procedures for deriving the optimal weights must be used when a test has more than three parts. First a derivation for the three-part case will be given. For the more than three-part case, a two-step derivation is given, and the two-step derivation will be shown to give the same result as the direct derivation in the three-part case.

A Derivation for the Three-Part Case

Using the assumptions given above about the congeneric model and without prior information about the congeneric coefficients, the reliability coefficient for a three-part congeneric test is given by (Kristof, 1974; Feldt & Brennan, 1989):

$$\rho_{XX'} \equiv \frac{\sigma_T^2}{\sigma_X^2} = \frac{\left(\sigma_{12}\sigma_{23} + \sigma_{12}\sigma_{13} + \sigma_{13}\sigma_{23}\right)^2}{\sigma_{12}\sigma_{23}\sigma_{13}\sigma_X^2} \quad , \tag{1}$$

where σ_{ij} is the covariance between part *i* and *j* and σ_i^2 is the variance of part *i*. Let

$$Y = w_1 X_1 + w_2 X_2 + w_3 X_3 \quad . \tag{2}$$

where the w_i's are weights. Usually, it is assumed that the weights sum to unity, but here for the convenience of a solution expressed below, the weights are not assumed to sum to one. Because multiplying the weights by a constant does not change the value of the reliability, the weights may be standardized to sum to one after they are found . The reliability of *Y* is

$$\rho_{YY'} = \frac{c\left(w_1\sigma_{12}\sigma_{13} + w_2\sigma_{12}\sigma_{23} + w_3\sigma_{13}\sigma_{23}\right)^2}{\left(w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + w_3^2\sigma_3^2 + 2w_1w_2\sigma_{12} + 2w_1w_3\sigma_{13} + 2w_2w_3\sigma_{23}\right)} \ , \tag{3}$$

where $c = 1/(\sigma_{12}\sigma_{13}\sigma_{23})$ is a constant which can be omitted in the process of finding the optimal weights.

One approach to solving this problem is to take the first partial derivatives with respect to w_1, w_2, and w_3. Then set the derivatives equal to zero and simultaneously solve for w_1, w_2, and w_3. To check if the results truly maximize the reliability, either the second derivatives can be taken and checked to see if they are negative at those solutions, or they can be empirically verified with some test data. It turns out the derivatives are too complicated to be presented in this paper. The equations will involve quadratic terms of the weights, which render them almost impossible to solve. But there is another indirect approach for solving this problem, and it is much simpler to present. This approach maximizes the numerator of the right side of Equation 3 while holding the denominator to be an arbitrary constant. If the relation between the weights is found not to depend on this constant, then this relation maximizes the reliability coefficient. As stated earlier, only the relation between the three weights, not the absolute value of the weights, affects the reliability coefficient. Maximizing this numerator is equivalent to maximizing the term inside the parenthesis because this term is positive. Using a Lagrange multiplier, the problem becomes one of finding the relative values of the weights that maximize

$$\begin{aligned} f = {} & w_2\sigma_{12}\sigma_{23} + w_1\sigma_{12}\sigma_{13} + w_3\sigma_{13}\sigma_{23} - \lambda(w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + w_3^2\sigma_3^2 \\ & +2w_1w_2\sigma_{12} + 2w_1w_3\sigma_{13} + 2w_2w_3\sigma_{23} - d)\ , \end{aligned} \tag{4}$$

where λ is the Lagrange multiplier (readers should not confuse this λ, which has no subscript, with the congeneric coefficients λ_i's, which have subscripts), and *d* is a constant at which the denominator of the Equation 3 is fixed.

Taking partial derivatives of f with respect to w_1, w_2, and w_3 and setting them to zero yields

$$\sigma_{12}\sigma_{13} - \lambda\left(2\sigma_1^2 w_1 + 2w_2\sigma_{12} + 2w_3\sigma_{13}\right) = 0 \tag{5}$$

$$\sigma_{12}\sigma_{23} - \lambda\left(2\sigma_2^2 w_2 + 2w_1\sigma_{12} + 2w_3\sigma_{23}\right) = 0 \tag{6}$$

$$\sigma_{13}\sigma_{23} - \lambda\left(2\sigma_3^2 w_3 + 2w_1\sigma_{13} + 2w_2\sigma_{23}\right) = 0\ . \tag{7}$$

Solving (5) for λ and substituting this expression into (6) and (7) yields the following two equations:

$$w_2 = \frac{\sigma_1^2\sigma_{23} - \sigma_{12}\sigma_{13}}{\sigma_2^2\sigma_{13} - \sigma_{12}\sigma_{23}} w_1 \tag{8}$$

$$w_3 = \frac{\sigma_1^2\sigma_{23} - \sigma_{12}\sigma_{13}}{\sigma_3^2\sigma_{12} - \sigma_{13}\sigma_{23}} w_1 \tag{9}$$

for the three unknowns w_1, w_2, and w_3. Note that this set of relationships does not depend on the constant *d*, which indicates that any set of weights that satisfies this set of relationships is optimal. Using *f*, *g*, and *h* as subscripts for the three parts (X_1, X_2, X_3) and letting f denote any one of the three parts, then the expression .

$$w_f = \left(\sigma_f^2\sigma_{gh} - \sigma_{fg}\sigma_{fh}\right)^{-1} , \tag{10}$$

satisfies Equations 8 and 9.

It can be seen that if the variances and covariances for all three parts are equal (which corresponds to the parts being parallel except for possible mean differences) , then equal weights are optimal. The formulas for optimal weights suggest that a part will get high weight if it has a small variance but large covariances with the other two parts.

Example One:

The following example used data from one form of the Work Keys Listening assessment (ACT, 1995). The data contain item scores for 1793 examinees who took six writing items. This example used the first three items in the test. Table 1 contains the results. The optimal weights are approximately .2, .6, and .2. The reliability increases from 0.618 for the unweighted sum to 0.677 for the weighted sum with the optimal weights. The increase is sizable, suggesting that the optimal weights might be useful in this case.

A Two-Step Derivation for the n-Part Case

Using the assumptions about the congeneric model given previously, and without prior information about the congeneric coefficients, the reliability coefficient for a test of four or more parts (assume there are *n* parts) is given by (Gilmer & Feldt, 1983; Feldt & Brennan, 1989):

$$\rho_{XX'} \equiv \frac{\sigma_T^2}{\sigma_X^2} = \left(\frac{\left(\sum D_f\right)^2}{\left(\sum D_f\right)^2 - \sum D_f^2}\right)\left(\frac{\sigma_X^2 - \sum \sigma_f^2}{\sigma_X^2}\right) , \tag{11}$$

where $D_f = \dfrac{\sum\limits_g \sigma_{fg} - \sigma_{fl} - \sigma_f^2}{\sum\limits_g \sigma_{lg} - \sigma_{fl} - \sigma_l^2}$, and row l is the row of the variance and covariance matrix with the largest sum of inter-part covariances. When $f = l$, $D_f = 1.0$.

This expression is so complicated that it is impossible to use the previous approach to find the optimal weights. A two-step approach to this problem is thus proposed here. In the first step, we will find a set of weights that makes the parts have equal true score variances, i.e., the parts become tau-equivalent. In the second step, we will find a set of optimal weights for the tau-equivalent model. The weights for each part thus obtained from these two steps are multiplied together to get the final weights for the original scores. It will be shown below that the weights thus obtained are optimal for the original part scores.

Theorem: For a test that has *n* parts with subscores $X_1, X_2, \cdots, X_n$, it is assumed that there exists a set of optimal (maximizing reliability) weights. If we transform X_i's to Y_i's by multiplying them with an arbitrary set of non-zero weights, $Y_i = w_i^{'} X_i$, i=1, 2,$\cdots$, n and if a set of weights $w_1^{''}, w_2^{''}, \cdots, w_n^{''}$ are optimal for $Y_1, Y_2, \cdots Y_n$. Then it is claimed that the set of weights $w_i = w_i^{'} w_i^{''}$, i=1, 2,$\cdots$, n , are optimal for $X_1, X_2, \cdots, X_n$.

Proof: By assumption, there is a set of optimal weights for $X_1, X_2, \cdots, X_n$. Denote them as $w_1^o, w_2^o, \cdots, w_n^o$, and let $\rho_{\max}$ be the maximum reliability.

Because $w_1^{''}, w_2^{''}, \cdots, w_n^{''}$ are optimal for $Y_1, Y_2, \cdots Y_n$, the reliability of the left side of the equation

$$w_1^{''} Y_1 + w_2^{''} Y_2 + \cdots + w_n^{''} Y_n = w_1^{''} w_1^{'} X_1 + w_2^{''} w_2^{'} X_2 + \cdots + w_n^{''} w_n^{'} X_n \qquad (12)$$

is also $\rho_{\max}$. Because if it is less than $\rho_{\max}$, then there exists another set weights $w_1^o / w_1^{'}, w_2^o / w_2^{'}, \cdots, w_n^o / w_n^{'}$, which when applied to $Y_1, Y_2, \cdots Y_n$ would make the reliability equal $\rho_{\max}$, and this contradicts the condition that $w_1^{''}, w_2^{''}, \cdots, w_n^{''}$ are optimal for

$Y_1, Y_2, \cdots Y_n$. If it is greater then $\rho_{\max}$, then by Equation 12, the right side of Equation 12 also has reliability greater than $\rho_{\max}$, which contradicts the condition that $\rho_{\max}$ is the maximum reliability of original parts using the optimal weights. Therefore, the left side of Equation 12 must have reliability equal to $\rho_{\max}$.

By Equation 12, we have that the weights $w_1, w_2, \cdots, w_n$ are optimal for $X_1, X_2, \cdots, X_n$., because the right side of Equation 12 also has reliability equal to $\rho_{\max}$. This completes the proof.

The two-step derivation for finding the optimal weights for an *n*-part test is described below.

Step One:

If we can estimate the λ_i's for each part, the inverse of the λ_i's are the weights that make the transformed part scores tau-equivalent. Thus, the step one weights are

$$w_i' = 1/\lambda_i \,, \quad i=1, 2, \cdots, n. \tag{13}$$

Now

$$Y = Y_1 + Y_2 + \cdots + Y_n, \tag{14}$$

where

$$\begin{aligned} Y_1 &= w_1' X_1 = (\lambda_1 T + E_1)/\lambda_1 = T + E_1/\lambda_1 \\ Y_2 &= w_2' X_2 = (\lambda_2 T + E_2)/\lambda_2 = T + E_2/\lambda_2 \\ &\vdots \\ Y_n &= w_n' X_n = (\lambda_n T + E_n)/\lambda_n = T + E_n/\lambda_n \end{aligned}$$

Step Two:

We need to find a set weights $w_1'', w_2'', \cdots, w_n''$ that are optimal for $Y_1, Y_2, \cdots Y_n$; that is, the reliability of $Z = w_1''Y_1 + w_2''Y_2 + \cdots + w_n''Y_n$ is maximized. To simplify the derivation, it is assumed that this set of weights sums to one (this condition is not necessary because if we fix the sum to be an arbitrary constant, we will have the same final solution). We have

$$Z = \left(w_1'' + \cdots + w_n''\right)T + \frac{w_1''}{\lambda_1}E_1 + \cdots + \frac{w_n''}{\lambda_n}E_n = T + \frac{w_1''}{\lambda_1}E_1 + \cdots + \frac{w_n''}{\lambda_n}E_n\ , \qquad (15)$$

$$\sigma_Z^2 = \sigma_T^2 + \frac{w_1''^2}{\lambda_1^2}\sigma_{E_1}^2 + \cdots + \frac{w_n''^2}{\lambda_n^2}\sigma_{E_n}^2\ . \qquad (16)$$

Notice that the true score part is not affected by the values of the weights. Hence the reliability may be maximized by minimizing the error variances in Equation 16. Again, using the Lagrange multiplier, the problem becomes one of minimizing the function *f* where

$$f = \frac{w_1''^2}{\lambda_1^2}\sigma_{E_1}^2 + \cdots + \frac{w_n''^2}{\lambda_n^2}\sigma_{E_n}^2 + \lambda\left(w_1'' + \cdots + w_n'' - 1\right)\ . \qquad (17)$$

Taking the first derivatives of *f* with respect to the w_i'''s, setting them to be zero and solving the resultant equations for w_i'''s yields

$$\frac{w_1''\sigma_{E_1}^2}{\lambda_1^2} = \frac{w_2''\sigma_{E_2}^2}{\lambda_2^2} = \cdots = \frac{w_n''\sigma_{E_n}^2}{\lambda_n^2} = -\frac{\lambda}{2}\ . \qquad (18)$$

By Equation 18, we have the relationships among the weights. Because only the relative values of the weights affect the reliability, any weights that satify these relationships are optimal. The following expression for the weights satisfies these relationships even though they do not add up to one.

$$w_i'' = \frac{\lambda_i^2}{\sigma_{E_i}^2}\ , \quad i=1,\ 2, \cdots, n\ , \qquad (19)$$

Then these $\ddot{w}_i$'s satisfy Equation 18, and combining them with Equations 13 and 19, yields the following final weights for the original scores:

$$w_i = \dot{w}_i \ddot{w}_i = \frac{1}{\lambda_i}\frac{\lambda_i^2}{\sigma_{E_i}^2} = \frac{\lambda_i}{\sigma_{E_i}^2}\ , \quad i\text{=}1,\ 2, \cdots, n. \tag{20}$$

The remaining problem is to find the error variances for the original score parts. Denoting the sum of the i'th row of the variance-covariance matrix of the original score parts as ψ_i, yields the following two equations

$$\psi_i = \sum_j \sigma_{ij} = \lambda_i(\lambda_1 + \lambda_2 + \cdots + \lambda_n)\sigma_T^2 + \sigma_{E_i}^2 = \lambda_i \sigma_T^2 + \sigma_{E_i}^2 \tag{21}$$

$$\sigma_i^2 = \lambda_i^2 \sigma_T^2 + \sigma_{E_i}^2 \tag{22}$$

Solving them yields

$$\sigma_{E_i}^2 = \frac{\sigma_i^2 - \lambda_i \psi_i}{1 - \lambda_i}\ . \tag{23}$$

Substituting the right side of Equation 23 into Equation 20, yields

$$w_i = \frac{\lambda_i(1 - \lambda_i)}{\sigma_i^2 - \lambda_i \psi_i}\ . \tag{24}$$

Noting that $\sum_{j \neq i} \sigma_{ij} = \psi_i - \sigma_i^2 = \lambda_i(1 - \lambda_i)\sigma_T^2$. Dropping the common term σ_T^2, we have

$$w_i = \frac{\psi_i - \sigma_i^2}{\sigma_i^2 - \lambda_i \psi_i}\ , \quad i\text{=}1,\ 2, \cdots, n\ . \tag{25}$$

Equations 24 and 25 are not equivalent, but they differ only by a constant σ_T^2, so they both represent the optimal weights.

To complete the solutions, formulas for the λ_i's are needed. Gilmer and Feldt (1983). provide the solutions for the λ_i's.

$$\lambda_i = \frac{D_i}{\sum_j D_j} \quad , \; i = 1, 2, \cdots, n. \tag{26}$$

All the computations for the weights are based on the variance-covariance matrix.

Example Two:

The following example used the same data as used in the previous example for the three parts. This example used all six items in the test. Table 2 contains the results of the computation. Item 5 gets the largest weight, followed by item 4 and item 6. The reliability increases from 0.751 for the unweighted sum to 0.796 for weighted sum with the optimal weights. This increase is moderate but is still valuable in this setting.

It is shown in the next part that Equation 25 gives the same results as Equation 10 in the three-part case. From Feldt and Brennan (1989), we have:

$$\lambda_f = \left(\frac{\sigma_{gh}}{\sigma_{fg}} + \frac{\sigma_{gh}}{\sigma_{fh}} + \frac{\sigma_{gh}}{\sigma_{gh}} \right)^{-1} . \tag{27}$$

Substituting this equation into Equation 25 gives

$$w_f = \frac{\sigma_{fg} + \sigma_{fh}}{\sigma_f^2 - \dfrac{\sigma_f^2 + \sigma_{fg} + \sigma_{fh}}{\dfrac{\sigma_{gh}}{\sigma_{fg}} + \dfrac{\sigma_{gh}}{\sigma_{fh}} + \dfrac{\sigma_{gh}}{\sigma_{gh}}}} = \frac{\sigma_{fg}\sigma_{fh} + \sigma_{fg}\sigma_{gh} + \sigma_{fh}\sigma_{gh}}{\sigma_f^2\sigma_{gh} - \sigma_{fg}\sigma_{fh}} . \tag{28}$$

The numerator of Equation 28 is the same for all the three weights, and thus can be dropped from the formula. It then gives the same expression as Equation 10.

Discussion

This paper gives two formulas for computing the weights that maximize test reliability, one for a three-part test, the other for a general case. It was shown that these two formula are consistent in the three-part case. There are some potential advantages for these formulas. First of all, they are easy to compute. Second, they enable us to gain insight into the factors that contribute to high or low weight for a particular part.

A natural question a reader might ask is about the two-part case. Because it is not possible to estimate the two congeneric coefficients based on one covariance, the approach presented in this paper cannot be applied to the two-part case. However, if the congeneric coefficients can be somehow obtained, then the general expression for the optimal weights in Equation 25 can still be applied to the two-part case.

As stated at the beginning of the paper, the decision on what weights to use may depend on a number of factors of which maximizing reliability may be just one. Wang and Stanley (1970) presented many different rationales for deriving the weights. Some of them are judgmental, others are empirically derived. The question of what factor should be weighted more than the others is entirely situation dependent. As reviewed by Wang and Stanley, however, using weights that maximizing the reliability are often considered a desirable alternative in the absence of an external criterion.

For most testing situations, particularly for those with high stakes on the part of the examinees, it is a good measurement practice to let the examinees know the score weighting at the testing time. So, it is necessary to collect pre-testing data for estimating the empirical weights such as the ones derived in this paper. It is usually not a good practice to change the weights after operational test administration. As with any sample of the data, the pre-test sample data collected for deriving these weights also contains sampling error. So the numbers computed using the those formulas should not taken at face value. It is advisable to estimate the weights based on more than one sample and compare the results whenever multiple samples are available.

A congeneric model is used in the derivation, which implies that if the situation is such that the congeneric model is not applicable, then these formula are probably also not applicable. How robust these formulas are to the deviation from the assumptions of the congeneric model needs to be studied empirically.

A final note is that these formula not only apply to performance assessment situations where they may be most useful, but that they also apply to other testing situations where each subtest may contain multiple items. They also apply to tests that contain both multiple-choice type items and constructed response type items. The major advantage of these formulas is that they only need the variance-covariance matrix of the part scores and do not need the reliability estimates of the part scores. In situations where reliability information for part scores are available, it is desirable to obtain weights using other procedures that require part score reliability estimates and compare them to weights derived using the formulas derived in this paper.

References

ACT (1995). *The Work Keys Assessment Program*. Iowa City, IA: ACT.

Bentler, P. M. (1968). Alpha-maximized factor analysis (Alphamax): Its relation to alpha and canonical factor analysis. *Psychometrika, 33*, 335-345.

Conger, A. J. (1974). Estimating profile reliability and maximally reliable composites. *Multivariate Behavioral Research, 9*, 85-104.

Feldt, L. S. & Brennan, R. L. (1989). Reliability in R. L. Linn (Ed.). *Educational Measurement* (3rd ed. pp.105-146). Washington D. C.: National Council on Measurement in Education and American Council on Education.

Gilmer, J. S. & Feldt, L. S. (1983). Reliability estimation for a test with parts of unknown lengths. *Psychometrika, 48*, 99-111.

Joe, G. W. & Woodward, J. A. (1976). Some developments in multivariate generalizability. *Psychometrika, 41*, 205-217.

Kaiser, H. F. & Caffrey, J. (1965). Alpha factor analysis. *Psychometrika, 30*, 1-14.

Kristof, W. (1974). Estimation of reliability and true score variance from a split of a test into three arbitrary parts. *Psychometrika*, 39, 491-499.

Li, H. (1997). A unifying expression for the maximum reliability of a linear composite. *Psychometrika, 62*, 245-249.

Li, H., Rosenthal, R., & Rubin, D. B. (1996). Reliability of measurement in psychology: from Spearman-Brown to maximal reliability. *Psychological Methods, 1*, 98-107.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Mosier, C. I. (1943). On the reliability of weighted composites. *Psychometrika, 8*, 161-168.

Peel, E. A. (1947). A short method for calculating maximum battery reliability. *Nature, London, 159*, 816-817.

Thomson, G. H. (1940). Weighting for battery reliability and prediction. *British Journal of Psychology, 30*, 357-366.

Wang, M. C., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research, 40*, 663-705.

Table 1. Computations for the three-part example.

Item	The variance-covariance matrix			Row sum	Cov. sum	Lambda	Weights
1	7.388	2.373	1.207	10.968	3.580	0.390	0.197
2	2.373	4.241	1.252	7.866	3.625	0.404	0.585
3	1.207	1.252	3.066	5.526	2.459	0.206	0.218

Table 2. Computations for the six-part example.

Item	The variance-covariance matrix						Row sum	Cov. sum	Lambda	weights
1	7.388	2.373	1.207	1.114	1.886	1.283	15.250	7.862	0.164	0.076
2	2.373	4.241	1.252	1.084	1.343	1.060	11.353	7.112	0.147	0.130
3	1.207	1.252	3.066	1.095	1.057	1.147	8.824	5.757	0.116	0.133
4	1.114	1.084	1.095	4.199	2.266	2.717	12.474	8.275	0.173	0.191
5	1.886	1.343	1.057	2.266	4.151	2.841	13.543	9.392	0.205	0.323
6	1.283	1.060	1.147	2.717	2.841	5.759	14.807	9.048	0.195	0.148

BEST COPY AVAILABLE

Reliability Issues and Possible Solutions

Catherine J. Welch
Dara R. Martinovich-Barhite

Abstract

Portfolio assessment is becoming popular in classrooms. ACT is striving to produce a portfolio assessment system that not only fosters good classroom instruction, but can be used to make decisions on a larger scale. Reliability is an important consideration in assessment, especially in assessment that is large-scale and high-stakes. This paper discusses the reliability of ACT's PASSPORT Portfolio System and how reliability has improved during PASSPORT's development. Possible reasons for this improvement in reliability, such as a broader distribution of scores and less variability due to readers, are examined.

Reliability Issues and Possible Solutions

Portfolio assessment is becoming a popular form of assessing student outcomes because it integrates classroom instruction and assessment. Many educators believe that portfolio assessment provides students the opportunity to use their classroom work, and their reflections on that work, to supply a richer and more valid picture of students' competencies than do other types of assessment (Gearhart & Herman, 1995). Portfolios challenge teachers and administrators to focus on meaningful outcomes, while providing a bridge between the worlds of public accountability and classroom practice. LeMahieu, Gitomer and Eresh (1995) assert that portfolio assessment supports instructional practice through the use of comprehensive and consistent tasks, providing detailed evidence of student thinking and encouraging students to become more active in their learning.

Portfolio assessment fits naturally with good instruction. In addition to its potential as a tool for thoughtful classroom assessment, portfolio assessment can also be used for large-scale testing (Freedman, 1993). The hope for connecting large-scale and classroom assessment is directly tied to gaining a better understanding of the measurement concerns and the classroom issues associated with the successful implementation of portfolios. ACT, in the development of a portfolio system, has attempted to address the practical issues, while striving to produce an equitable, technically-sound assessment system. One of the more important issues associated with the development of a portfolio system is the ability to consistently evaluate student responses. Consistency (or, *reliability*) is desirable, because it demonstrates that scoring criteria are being applied fairly, which means that everyone can agree on the significance of the results. Reliable results may be interpreted, summarized, compared, and used to make important decisions.

The purpose of this paper is to document the reliability results of ACT's PASSPORT portfolio system following year one, the changes that were instituted between years one and two to address reliability concerns, and the results of the year two pilot after the changes had been incorporated.

An Overview of PASSPORT

ACT's portfolio system, PASSPORT, is an assessment system designed to augment existing assessments, improve instruction, and complement ongoing classroom activities. The skills assessed by PASSPORT have been selected from an analysis of national curriculum standards and state curriculum frameworks, and PASSPORT provides a means for directly connecting class activities to these standards and frameworks. The portfolio involves students from start to finish and reflects real class work. Students and teachers enter into a dialogue as they agree on goals, define projects, and assess progress. Although PASSPORT is classroom-centered and flexible enough to adapt to a variety of school settings, PASSPORT also provides a framework of materials and support services.

ACT's portfolio system is the product of a three-year developmental process. During the first year of the project, ACT staff worked with schools (secondary and post-secondary), employers, national educational organizations, and curriculum organizations to identify the potential need for and uses of a large-scale portfolio assessment system. The goal for the first year of the project was to determine the feasibility of developing a portfolio system that could meet a wide variety of needs for a wide variety of groups and to identify what materials and services should be provided by such a system. The second year of development involved the field test of PASSPORT with a sample of seven schools that represented a range of school types, geographic locations, and student populations. These schools were selected to be as representative as possible of the nation's diverse geography and population. An initial pilot administration followed the field test during the third year of development; at that point, the Work Sample Descriptions and accompanying scoring rubrics were prepared for a second round of piloting. It is the changes seen between the field test administration (1994-95) and the pilot administration (1995-96) that are summarized in this paper.

Work Sample Descriptions

So that it may fit within a classroom setting, PASSPORT provides flexible guidelines concerning the types of student work that can be submitted as portfolio entries. The portfolio framework is built from a menu of broadly-defined categories of activities that students already do in class. These categories are called Work Sample Descriptions. For each Work Sample Description, a variety of class activities can produce related student work. Teachers select the five most appropriate Work Sample Descriptions from this menu of options for assignment in their particular classes. Students are asked to submit work from their regular class assignments that matches each of these five different Work Sample Descriptions. Each student is provided an opportunity to complete, select, and reflect upon the work that will represent their skills in each of three areas: Language Arts, Mathematics, and Science. Thus, students complete five entries in a particular content area.

The PASSPORT Science portfolio component is specifically designed to cut across a broad range of skills that are of value in the secondary science classroom. A variety of approaches is presented to the teacher and student that allows for the identification and selection of student work that matches class activities. For example, the Work Sample Descriptions in Science include literature review and evaluation, integrating sciences, the societal context of science, historical perspectives, evaluating scientific claims, laboratory observations, laboratory experiments, designing studies, and performing studies. Each of these Work Sample Descriptions is evaluated using its own rubric, which lists criteria relevant to the task being performed. These criteria evaluate the student's proficiency at communicating the depth of scientific understanding, specifying the appropriate purpose and hypotheses, developing and following an appropriate design, presenting procedures and results in an organized and appropriate format, analyzing and evaluating information, drawing conclusions, and using and citing varied sources of information.

The PASSPORT Mathematics portfolio component provides the opportunity to analyze data, use mathematics to solve problems from another class, solve challenging problems, collect and analyze data, compare notions, use a technological tool, construct logical arguments, solve a problem using multiple solution strategies, solve real-world problems, and show connections among branches of mathematics. The Mathematics component also contains a different rubric for each Work Sample Description. Some of the features used to evaluate student work are the choice of problem, description of the problem, accuracy of analysis, correctness and interpretation of data, correctness of solution, interpretation of results, justification, understanding, and comparison of concepts.

The PASSPORT Language Arts portfolio follows the same format as the Science and Mathematics in that it encompasses a broad range of activities. The Language Arts Work Sample Descriptions allow teachers to select from explanation, analysis and evaluation, business and technical writing, poetry, writing a short story or drama, persuasive writing, relating a personal experience, research/investigative writing, responding to a literary text, writing a review of the arts or media, and writing about the uses of language. Each Work Sample Description is evaluated according to its own rubric. Some of the common features found in these rubrics include completeness, development, clarity, audience awareness, voice, word choice, sentence variety and mechanics.

Finally, PASSPORT requires students to reflect on their learning and accomplishments by writing a self-reflective cover letter. One of the most important benefits of PASSPORT is a student's self-reflection on his or her growth and development as a learner. The cover letter is intended to help people who read the student's portfolio understand how the portfolio demonstrates the student's mastery of specific skills and concepts and how that mastery relates to the student's growth and goals.

Scoring

A modified holistic scoring procedure was adopted for the scoring of PASSPORT results. Each entry (there are five entries per content area) receives a single score on a *six-point* scale. In addition, the entire student portfolio receives an overall score, on a *four-point* scale, that takes into account the features found at the individual Work Sample Description level as well as the variety of entries and evidence of growth and depth found in the self-reflective letter and the entries. This paper will focus on the individual Work Sample Description results, on the six-point scale.

During the development of PASSPORT, a specific scoring rubric was designed for each Work Sample Description, and actual student responses from the pilot test administration were used to illustrate each score point of the rubric. Teachers and ACT staff who participated in this rubric-writing process examined student responses from all participating schools, taking into account the varied interpretations and approaches to the particular Work Sample Descriptions across the schools. This review process helped to ensure that various cultural backgrounds, course offerings, and opportunities were taken into account. Based on this process, the Work

Sample Descriptions were refined to be as broad as possible while still considering readers' ability to evaluate them in a consistent manner. Readers noted particular difficulties associated with Work Sample Descriptions during the scoring process. This information was used to further refine the Work Sample Descriptions. Work Sample Descriptions that proved to be too difficult or were misinterpreted in their intent were reviewed and revised prior to the second pilot test administration.

As with the development and design of the Work Sample Descriptions, a variety of classroom teachers, multicultural educators, content experts, and measurement specialists worked to develop the scoring rubrics. Throughout the training and scoring process, reader consistency was monitored, evaluated, and documented.

Reliability

The reliability of the portfolio was addressed by two separate analyses. To address reliability, 25% of the portfolios, sampled randomly, were evaluated by a second reader. The first analysis estimates indices of reader agreement [such as interrater reliability (Pearson's correlations) and interrater agreements (expressed in percents)], which describe the degree to which readers agree with each other when scoring the same work sample. Indices of reader agreement identify how well the scoring standards have remained fixed throughout the scoring process. Interrater agreements serve as an indication of the degree to which the responses have required a third reading, as the percent of papers requiring a third reading is expressed as the percent of papers whose scores were *resolved*. The third readers were team leaders, who supervised the readers and who had more experience working on scoring projects.

The second reliability analysis, called generalizability analysis, is used to estimate the various sources of measurement error. The scoring process is designed so that the most appropriate variance sources (such as the particular Work Sample Descriptions chosen, the sample of examinees, readers, and various interaction components) can be identified and estimated. A reliability-like coefficient, the generalizability coefficient, is estimated from this analysis. For these generalizability analyses, the SAS procedure MIVQUE was used in both 1994-95 and 1995-96 to estimate variance components and assess generalizability. For a more thorough discussion of generalizability designs and analyses, see the Gao and Colton (1997) paper in this report.

Results

During the 1994-1995 academic year, teachers at seven field test sites participated in the project. During the 1995-1996 academic year, teachers at 20 pilot sites used PASSPORT in their classrooms. At the end of each school year, students compiled their work into finished portfolios which were sent to ACT for scoring. Scores were assigned on a scale of 1 to 6 for each Work Sample Description and on a scale of 1 to 4 for the overall portfolio. Readers were content-area experts, and most had teaching experience in the secondary classroom. Readers were trained to score according to ACT-developed rubrics and needed to qualify before scoring began.

Tables 1, 2, and 3 (at the end of this report) show the descriptive statistics and frequency distributions for each of the Language Arts, Mathematics, and Science Work Sample Descriptions for the 1994-95 and 1995-1996 academic years. Italics denote 1994-95 data, which are recorded below 1995-96 data.

Because the group assessed in 1995-96 was different than the group in 1994-95 (although both groups were selected because they represented the entire spectrum of the national educational system), differences in means and frequencies might have been due to these group differences. However, at least some of the differences were due to teachers' having more experience with PASSPORT and to changes made to the PASSPORT system.

Language Arts

Overall, the results for the Language Arts portfolio were consistent from year one to year two. Year two showed an overall increase in the mean performance on Work Sample Descriptions, a decrease in the percent of low scores that were assigned and a slight increase in the percent of high scores that were assigned. In 1994-95, the highest mean scores were obtained on Analysis/Evaluation (3.17), Relating a Personal Experience (3.14), and Explanatory Writing (3.15). In 1995-96, the highest mean scores on the individual Work Sample Descriptions were obtained on the Research/Investigative Writing (3.47) and Analysis/Evaluation (3.46) Work Sample Descriptions.

Means should not be interpreted without looking at the standard deviations and frequency distributions. In 1995-96, the standard deviations of scores on the individual Work Sample Descriptions ranged from 0.95 to 1.22, which shows that scores tended to cluster within a score point or so of each mean. These were fairly consistent with 1994-95 results that ranged from .88 to 1.30. However, overall there was a slight decrease in the standard deviations.

The frequency distributions in Table 1 also show where scores tend to cluster within each Work Sample Description. In both years, most language arts Work Sample Descriptions showed more scores in the lower-score end of the distribution than in the upper-end.

Mathematics

Comparing the individual Work Sample Descriptions (rated on a scale of 1 to 6), the highest mean scores in 1995-96 were obtained on the Logical Argument (3.23) and Challenging Problem (3.18) Work Sample Descriptions. The highest means in 1994-95 were found for Logical Argument (4.30), Another Class (3.92), and Technology (3.88). In 1994-95 the mean performance on Work Sample Descriptions ranged from 1.63 to 4.30, and in 1995-96 the means ranged from 2.22 to 3.23. The change in performance between years one and two was not as systematic as found with the Language Arts. In Mathematics, only four of the 11 Work Sample Descriptions showed increases in mean scores from 1994-95 to 1995-96. The rest showed decreases. This may be due to the reworking of some Work Sample Descriptions and their rubrics so that scores were distributed more evenly, making it harder to receive top scores. The fact that more significant changes were seen between years in Mathematics may also be due to an increase in the number of participating teachers and the variety of mathematics classes that were included in the second year of the pilot. In addition, small sample sizes during the first year likely contributed to unstable estimates of performance.

Unlike the frequency distributions in Language Arts portfolios, the distribution of scores on some of the Mathematics Work Sample Descriptions are frequently bimodal. For example, in 1995-96, the scores on the From Your Own Experience Work Sample Description peaked at a score of 1 and at a score of 3. Scores on the Logical Argument Work Sample Description show a large peak at a score of 3 and a smaller peak at a score of 5. Technology scores demonstrate a large peak at 3 and a smaller peak at 1. A bimodal distribution could mean that the Work Sample Description tended to be chosen in higher- and lower-level classes or the curriculum emphasizes the necessary skills in higher and lower grades. Within all Mathematics Work Sample Descriptions, there are more scores in the lower end of the distribution than in the upper end. This was true for both years.

The standard deviations on the individual Work Sample Descriptions ranged from 0.85 to 1.54 in 1995-96. These results were similar for 1994-95 when the standard deviations ranged from .64 to 1.54. In 1995-96, the largest standard deviation, 1.54, was seen in the scores of the

Logical Argument Work Sample Description, which had a large peak of scores at 3 and a smaller peak at 5 in 1996.

Science

In 1995-95, the individual Work Sample Descriptions receiving the highest average scores were Literature Review and Evaluation (2.63) and Historical Perspective (2.44). Students in 1994-95 had also performed best on these same Work Sample Descriptions, with means of 2.64 for Literature Review and 1.95 for Historical Perspective.

In Science, none of the average Work Sample Description scores for 1994-95 or 1995-96 was 3 or over. Scores of 5 and 6 were more infrequent in science than in either Mathematics or Language Arts. All of the science distributions had one peak, situated closer to the lower end of the score scale.

The standard deviations of scores on the individual Work Sample Descriptions ranged from 0.49 to 1.01 in 1994-95 and from .69 to 1.03 in 1995-96. Scores in Science clustered more tightly than scores in Mathematics and Language Arts, as evidenced by fewer scores at the higher end of the score scale in Science.

Reliability Results

In both years, twenty-five percent of the PASSPORT portfolios were double-scored by a randomly-selected second reader to provide estimates of reliability. Indices of interrater reliability [*interrater correlations* (Pearson's) and the percentage of scores in the *perfect agreement*, *adjacent agreement*, and *resolved* categories] were computed for each Work Sample Description. *Perfect agreement* was achieved when both readers assigned the same score to the student's entry. *Adjacent agreement* was achieved when the two scores assigned to the student's entry were within one point of each other. *Resolved* scores were originally more than one point apart and were settled through discussion among the two readers and the team leader (who serves as the third reader).

Tables 4, 5, and 6 show the interrater statistics and accuracy statistics for each Language Arts, Mathematics and Science Work Sample Description in each content area for both years. The sample size was larger in 1995-96 than it was in 1994-95. Tables 4, 5, and 6 provide indices of interrater reliability for only the Work Sample Descriptions for which 25 or more papers were double-scored. Results from 1994-95 are in italics, and results from 1995-96 are in plain text.

Language Arts

In 1995-96, in Language Arts, the percentage of readers in perfect agreement ranged from 73.9% for Evaluation of Print or Electronic Media to 49.2% for Proposing a Solution. The percentage of readers in perfect or adjacent agreement ranged from 100% for Business and Technical Writing and Proposing a Solution to 96% for Imaginative Writing.

As can be seen in Table 4, even the Work Sample Descriptions with the lowest interrater agreements in 1995-96 still demonstrate good agreement among readers. This shows that readers were in solid agreement with each other, most likely due to adherence to rubrics. In 1994-95, the percentage of readers in perfect agreement ranged from 60.7% for Writing about Values, Issues, and Beliefs to 34.3% for Persuasive Writing.

In 1994-95, the median interrater correlation was .60 with a high of .79 and a low of .47. The median interrater correlation in 1995-96 in Language Arts was 0.78, with a low of 0.68 for Writing about Uses of Language to a high of 0.86 for Business and Technical Writing. These are all moderate to high correlations for portfolio assessment. Changes (described later) between the two years seemed to result in substantial increases in interrater correlations.

Mathematics

Among Mathematics Work Sample Descriptions in 1995-96, the percentage of readers in perfect agreement ranged from 84.0% for Connections to 51.0% for Logical Argument. In 1994-95, the percent in perfect agreement ranged from 42.5% to 90.6%. There was an overall increase in the accuracy of the readers between the two years.

In 1995-96, the percentage of readers in perfect or adjacent agreement ranged from 99.3% for Connections to 82.0% for Logical Argument. For the 1995-96 Logical Argument Work Sample Description, 18.0% of readers' scores were resolved. All of the rest of the 1995-96 Mathematics Work Sample Descriptions had 6.0% or fewer of their readers falling into this category. These results may be compared to 1994-95 results showing perfect or adjacent agreement of 100% for Collecting and Analyzing Data to 79.9% for Comparing Notions. In 1994-95, Comparing Notions had the largest percentage of papers needing resolution, at 20.1%. Logical Argument had about the same percentage of papers needing resolution, at 17.1%.

In 1995-96, the median interrater correlation in mathematics was 0.79, with a low of 0.58 for Logical Argument and a high of 0.89 for Connections. The interrater correlations for all of the Work Sample Descriptions were 0.70 or higher, except for Logical Argument. Similar statistics in 1994-95 ranged from 0.46 for Comparing Notions to 0.96 for Collecting/Analyzing Data. The change between 1994-95 and 1995-96 was not as consistent as with Language Arts. The interrater correlations tended to fluctuate in both directions. Small, unstable samples in 1994-95 likely contributed to artificially high estimates.

Science

For the individual Work Sample Descriptions in 1995-96, the percentage of readers in perfect agreement ranged from 82.7% for Design and Perform a Study to 69.0% for Literature

Review and Evaluation. Also in 1995-96, the percentage of readers in perfect or adjacent agreement ranged from 100% in Applications, Design and Perform a Study, Historical Perspective, and Literature Review and Evaluation to 99.0% in Laboratory Experiment. Science had very high interrater agreement in 1995-96, most likely due to strict adherence to rubrics.

In 1995-96, the median interrater correlation among Science Work Sample Descriptions was 0.77, with a low of 0.72 for Design a Study and a high of 0.86 for Applications. In 1994-95, similar values ranged from 0.44 to 0.62. Overall there was a positive effect on interrater reliability between the two years. All Work Sample Descriptions increased with respect to the interrater correlation and decreased with respect to the percent of papers needing resolution.

Generalizability Results

As seen in Table 7, the generalizability coefficients for the 1995-96 pilot were Language Arts (0.75), Mathematics (0.79), and Science (0.65). These represented changes from the 1994-95 year of Language Arts (0.73), Mathematics (0.33) and Science (0.31). Values within the 1995-96 range are expected, given the number of work samples a student submits (five, which is far fewer than the number of items found on a typical multiple-choice test) and the fact that human judgment is used in scoring, even though rubrics keep scoring as objective as possible. The Discussion section seeks to explain the differences in generalizability between the two years, especially in Mathematics and Science.

Discussion

The second year results, in all three content areas, were overall more reliable than the first year results. However, Mathematics did have two exceptions to this generalization. The increase in reliability is likely due to two factors: a broader distribution of scores that represented the entire score range and a decrease in the variability due to readers. These effects were the result of a number of changes that were instituted between years. These changes were deliberately introduced into the program following a review of the first year results. The changes were also introduced and implemented in a larger sample of classrooms than the initial framework. These changes included:

1. Work Sample Descriptions were more structured during year two than they were during year one. This additional structure helped participating teachers to focus on assignments that were appropriate for each Work Sample Description. The teachers and students spent more time in the selection of the appropriate sample of student work than they did the first year.

2. Teachers were given more examples of student work at each of the score points than they were the first year. Prior to the beginning of the academic year, samples of student work were shared with the teachers during an initial staff development workshop. Additional samples of student work were shared with teachers midway through the academic year.

3. More examples of classroom assignments were provided to participating teachers. These assignments were selected from those that were submitted the first year and may have provided more of a context for teachers who were selecting activities for the work samples.

4. Teachers participated in a two-day workshop that provided an exposure to the scoring criteria and scoring practices used by ACT. This workshop provided teachers with a variety of examples of student work and articulations for the assigned scores. ACT staff worked with participating teachers to become more familiar with the scoring criteria during this workshop. Teachers attending the workshop had the opportunity to evaluate student work with the scoring guides that accompany the program.

5. Scoring criteria were shared at the beginning of the school year with students and teachers. This early dissemination of information helped both students and teachers to focus on the evaluative criteria throughout the entire year.

6. Practicing teachers were hired as readers and trained by ACT staff to internalize the scoring rubrics. The selection of practicing teachers helped to address the issue of expectations and helped to define the scale used by the readers.

7. Readers were trained specific to each Work Sample Description in both years one and two. However, year two readers were provided with more clear examples of what type of performance constituted each of the possible points on the score scale.

Conclusions

The successful implementation of a portfolio system that includes an assessment component must include a refined set of rubrics that have been field-tested and pilot-tested on a representative group of students. The field test and pilot test must be designed to collect not only student information but also information from the teachers with regard to impact, correspondence to curriculum, interpretability, and generalizability.

Reliability of portfolio results can be increased through the systematic exposure of the scoring rubric and assignments to participating teachers. Students must also know and be able to understand the scoring rubric and the tie between examples of work selected for inclusion in the portfolio and the scoring process.

In a large-scale assessment environment, there must be some constraints placed on the types of assignments and selection of student work to enhance the ability to evaluate the work reliably. A system that allows for student selection without these guidelines and constraints will lead to results that are not generalizable beyond the specific assignment.

References

Gearhart, M. & Herman, J. (1995). *Portfolio Assessment: Whose Work Is It? Issues in the Use of Classroom Assignments for Accountability.* (CSE Tech Rep). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Freedman, S.W. (1993). Linking large-scale testing and classroom portfolio assessments of student writing. *Educational Assessment, 1,* 27-52.

LeMahieu, P., Gitomer, D. & Eresh, J. (1995). *Portfolios beyond the classroom: Data quality and qualities.* Princeton, NJ: Center for Performance Assessment, Educational Testing Service.

TABLE 1
Distribution of Language Arts Scores
(Note: 1995-96 results are in plain text; *1994-95 results are in italics*)

Work Sample Description	Distribution by Score Level (%)	Mean (out of 6)	S.D.	Number
Analysis/Evaluation	1 = 2.0% *7.9%*	3.46 *3.17*	0.95 *1.13*	1325 *313*
	2 = 10.3% *19.1%*			
	3 = 42.6% *35.0%*			
	4 = 32.1% *27.1%*			
	5 = 11.3% *9.4%*			
	6 = 1.7% *1.5%*			
Business and Technical Writing	1 = 9.4% *27.8%*	2.95 *2.14*	1.10 *0.88*	235 *71*
	2 = 24.7% *40.0%*			
	3 = 36.2% *26.7%*			
	4 = 22.1% *5.6%*			
	5 = 6.4% *0%*			
	6 = 1.3% *0%*			
Evaluation of Print or Electronic Media	1 = 8.7% *25.0%*	2.95 *2.58*	0.99 *1.21*	219 *96*
	2 = 18.3% *21.0%*			
	3 = 47.5% *32.0%*			
	4 = 21.5% *15.0%*			
	5 = 2.7% *7.0%*			
	6 = 1.4% *0%*			

TABLE 1
Distribution of Language Arts Scores
(Note: 1995-96 results are in plain text; *1994-95 results are in italics*)

Work Sample Description	Distribution by Score Level (%)	Mean (out of 6)	S.D.	Number
Explanatory Writing	1 = 4.1% *5.8%*	3.17 *3.15*	1.00 *1.13*	972 *67*
	2 = 19.7% *27.5%*			
	3 = 41.7% *30.4%*			
	4 = 25.7% *21.7%*			
	5 = 7.6% *14.5%*			
	6 = 1.2% *0%*			
Imaginative Writing	1 = 5.4% *15.8%*	3.11 *2.64*	1.09 *1.18*	1960 *213*
	2 = 25.4% *35.0%*			
	3 = 34.6% *29.1%*			
	4 = 24.0% *12.8%*			
	5 = 9.4% *6.0%*			
	6 = 1.2% *1.3%*			
Persuasive Writing	1 = 1.2% *17.2%*	3.35 *2.95*	0.96 *1.30*	1371 *291*
	2 = 16.3% *16.9%*			
	3 = 42.3% *32.4%*			
	4 = 28.1% *20.6%*			
	5 = 10.6% *10.5%*			
	6 = 1.5% *2.4%*			

BEST COPY AVAILABLE

TABLE 1
Distribution of Language Arts Scores
(Note: 1995-96 results are in plain text; *1994-95 results are in italics*)

Work Sample Description	Distribution by Score Level (%)	Mean (out of 6)	S.D.	Number
Proposing a Solution	1 = 8.0%	2.91	1.04	261
	8.7%	*2.92*	*1.09*	*126*
	2 = 25.7%			
	29.1%			
	3 = 40.6%			
	32.3%			
	4 = 19.5%			
	22.0%			
	5 = 5.0%			
	7.9%			
	6 = 1.1%			
	0%			
Relating a Personal Experience	1 = 1.7%	3.32	0.98	2093
	11.9%	*3.14*	*1.24*	*342*
	2 = 17.3%			
	16.9%			
	3 = 41.1%			
	30.7%			
	4 = 27.8%			
	25.8%			
	5 = 10.8%			
	13.0%			
	6 = 1.2%			
	1.7%			
Research/Investigative Writing	1 = 4.5%	3.47	1.22	1314
	19.3%	*2.77*	*1.24*	*203*
	2 = 18.6%			
	22.4%			
	3 = 27.4%			
	28.1%			
	4 = 28.2%			
	22.8%			
	5 = 17.3%			
	6.6%			
	6 = 4.0%			
	0.9%			

TABLE 1
Distribution of Language Arts Scores
(Note: 1995-96 results are in plain text; *1994-95 results are in italics*)

Work Sample Description	Distribution by Score Level (%)	Mean (out of 6)	S.D.	Number
Responding to a Literary Text	1 = 7.9% *14.1%*	3.02 *2.72*	1.11 *1.10*	2194 *278*
	2 = 25.8% *28.5%*			
	3 = 32.7% *32.4%*			
	4 = 24.6% *19.7%*			
	5 = 8.2% *4.6%*			
	6 = 0.8% *0.7%*			
Writing about an Out-of-Class Reading	1 = 8.9% *9.5%*	3.01 *2.81*	1.17 *1.11*	640 *35*
	2 = 26.9% *35.7%*			
	3 = 31.3% *28.6%*			
	4 = 22.5% *21.4%*			
	5 = 8.6% *2.4%*			
	6 = 1.9% *2.4%*			
Writing about Uses of Language	1 = 0% **NA*	3.72 **NA*	0.94 **NA*	86 **NA*
	2 = 9.3%			
	3 = 32.6%			
	4 = 36.0%			
	5 = 20.9%			
	6 = 1.2%			

BEST COPY AVAILABLE

TABLE 1
Distribution of Language Arts Scores
(Note: 1995-96 results are in plain text; *1994-95 results are in italics*)

Work Sample Description	Distribution by Score Level (%)	Mean (out of 6)	S.D.	Number
Writing about Values/Issues/Beliefs	1 = 3.7% *21.5%*	3.10 *2.78*	1.00 *1.28*	1051 *131*
	2 = 23.9% *27.1%*			
	3 = 40.9% *25.7%*			
	4 = 22.5% *16.0%*			
	5 = 8.1% *7.6%*			
	6 = 0.9% *2.1%*			
Writing a Review of the Visual or Performing Arts	1 = 6.5% *8.3%*	2.94 *3.64*	1.05 *1.12*	307 *11*
	2 = 30.3% *33.3%*			
	3 = 33.2% *41.7%*			
	4 = 22.8% *16.7%*			
	5 = 6.8% *0%*			
	6 = 0.3% *0%*			
Overall Portfolio Score	**Distribution by Score Level (%)**	**Mean (out of 4)**	**S.D.**	**Number**
Overall Language Arts Portfolio	1 = 32.3%	1.93 *2.20*	0.79 *0.90*	3341 *619*
	2 = 45.1%			
	3 = 19.8%			
	4 = 2.8%			

** In 1994-95, there were so few entries corresponding to this Work Sample Description that meaningful indices could not be obtained.*

TABLE 2
Distribution of Mathematics Scores
(Note: 1995-96 results are in plain text; *1994-95 results are in italics*)

Work Sample Description	Distribution by Score Level (%)	Mean (out of 6)	S.D.	Number
Another Class	1 = 11.4% *0%*	3.02 *3.92*	1.05 *0.64*	458 *13*
	2 = 13.1% *7.7%*			
	3 = 43.2% *0%*			
	4 = 27.7% *84.6%*			
	5 = 3.5% *7.7%*			
	6 = 1.1% *0%*			
Analyzing Data	1 = 28.5% *25.0%*	2.22 *2.55*	1.06 *1.24*	810 *60*
	2 = 36.9% *23.3%*			
	3 = 21.1% *31.7%*			
	4 = 11.2% *13.3%*			
	5 = 1.7% *5.0%*			
	6 = 0.5% *1.7%*			

BEST COPY AVAILABLE

(table 2 cont.)

TABLE 2
Distribution of Mathematics Scores
(Note: 1995-96 results are in plain text; *1994-95 results are in italics*)

Work Sample Description	Distribution by Score Level (%)	Mean (out of 6)	S.D.	Number
Connections	1 = 22.0% *47.9%* 2 = 47.1% *42.5%* 3 = 18.8% *8.2%* 4 = 7.3% *1.4%* 5 = 4.3% *0%* 6 = 0.6% *0%*	2.26 *1.63*	1.05 *0.70*	537 *73*
Collecting and Analyzing Data	1 = 21.0% *62.0%* 2 = 40.5% *13.9%* 3 = 32.8% *17.6%* 4 = 5.4% *4.6%* 5 = 0.3% *1.9%* 6 = 0% *0%*	2.23 *1.70*	0.85 *1.03*	671 *108*

TABLE 2
Distribution of Mathematics Scores
(Note: 1995-96 results are in plain text; *1994-95 results are in italics*)

Work Sample Description	Distribution by Score Level (%)	Mean (out of 6)	S.D.	Number
Consumer Beware	1 = 33.2% *50.0%*	2.39 *1.88*	1.15 *1.20*	334 *16*
	2 = 14.1% *31.3%*			
	3 = 34.7% *6.3%*			
	4 = 17.1% *6.3%*			
	5 = 0.6% *6.3%*			
	6 = 0.3% *0%*			
Comparing Notions	1 = 18.3% *7.4%*	2.45 *3.74*	1.04 *1.31*	273 *108*
	2 = 37.4% *10.2%*			
	3 = 27.8% *22.2%*			
	4 = 13.9% *25.0%*			
	5 = 2.2% *31.5%*			
	6 = 0.4% *3.7%*			

BEST COPY AVAILABLE

(table 2 cont.)

TABLE 2
Distribution of Mathematics Scores

(Note: 1995-96 results are in plain text; *1994-95 results are in italics*)

Work Sample Description	Distribution by Score Level (%)	Mean (out of 6)	S.D.	Number
Challenging Problem	1 = 6.9% *13.3%*	3.18 *2.82*	1.09 *0.91*	1237 *442*
	2 = 15.0% *6.3%*			
	3 = 45.9% *70.6%*			
	4 = 19.2% *4.8%*			
	5 = 11.5% *4.5%*			
	6 = 1.5% *0.5%*			
From Your Own Experience	1 = 34.8% *9.5%*	2.17 *2.95*	1.02 *0.89*	279 *347*
	2 = 21.9% *4.3%*			
	3 = 35.8% *75.2%*			
	4 = 6.5% *5.5%*			
	5 = 1.1% *3.7%*			
	6 = 0% *1.7%*			

(table 2 cont.)

TABLE 2
Distribution of Mathematics Scores
(Note: 1995-96 results are in plain text; *1994-95 results are in italics*)

Work Sample Description	Distribution by Score Level (%)	Mean (out of 6)	S.D.	Number
Logical Argument	1 = 15.6% *8.7%*	3.23 *4.30*	1.54 *1.54*	409 *149*
	2 = 18.1% *3.4%*			
	3 = 28.9% *16.8%*			
	4 = 10.5% *19.5%*			
	5 = 18.3% *23.5%*			
	6 = 8.6% *28.2%*			
Multiple Methods	1 = 22.4% *16.2%*	2.55 *2.80*	1.14 *1.14*	939 *358*
	2 = 22.7% *19.0%*			
	3 = 38.6% *39.9%*			
	4 = 11.5% *19.8%*			
	5 = 3.7% *4.7%*			
	6 = 1.2% *0.3%*			

BEST COPY AVAILABLE

(table 2 cont.)

TABLE 2
Distribution of Mathematics Scores
(Note: 1995-96 results are in plain text; *1994-95 results are in italics*)

Work Sample Description	**Distribution by Score Level (%)**	**Mean (out of 6)**	**S.D.**	**Number**
Technology	1 = 20.3% *1.9%* 2 = 13.0% *3.8%* 3 = 36.7% *41.5%* 4 = 19.4% *10.8%* 5 = 9.4% *41.5%* 6 = 1.1% *0.5%*	2.88 *3.88*	1.26 *1.08*	561 *212*
Overall Portfolio Score	**Distribution by Score Level (%)**	**Mean (out of 4)**	**S.D.**	**Number**
Overall Mathematics Portfolio	1 = 35.7% 2 = 52.1% 3 = 12.1% 4 = 0.1%	1.77 *2.11*	0.65 *0.81*	1588 *538*

TABLE 3
Distribution of Science Scores
(Note: 1995-96 results are in plain text; *1994-95 results are in italics*)

Work Sample Description	Distribution by Score Level (%)	Mean (out of 6)	S.D.	Number
Applications	1 = 30.6%	2.09	0.91	722
	66.0%	*1.46*	*0.73*	*153*
	2 = 36.4%			
	23.5%			
	3 = 27.3%			
	8.5%			
	4 = 5.1%			
	2.0%			
	5 = 0.6%			
	0%			
	6 = 0%			
	0%			
Design a Study	1 = 46.5%	1.65	0.69	396
	*NA***	*NA***	*NA***	*NA***
	2 = 42.4%			
	3 = 10.4%			
	4 = 0.8%			
	5 = 0%			
	6 = 0%			
Design and Perform a Study	1 = 31.3%	1.88	0.74	275
	33.7%	*1.92*	*0.83*	*89*
	2 = 51.3%			
	44.9%			
	3 = 15.6%			
	16.9%			
	4 = 1.5%			
	4.5%			
	5 = 0.4%			
	0%			
	6 = 0%			
	0%			

BEST COPY AVAILABLE

(table 3 cont.)

TABLE 3
Distribution of Science Scores
(Note: 1995-96 results are in plain text; *1994-95 results are in italics*)

Work Sample Description	Distribution by Score Level (%)	Mean (out of 6)	S.D.	Number
Evaluating Scientific Claims	1 = 31.5%	1.88	0.74	653
	75.4%	*1.27*	*0.49*	*191*
	2 = 51.9%			
	22.5%			
	3 = 13.9%			
	2.1%			
	4 = 2.6%			
	0%			
	5 = 0%			
	0%			
	6 = 0%			
	0%			
Historical Perspective	1 = 7.7%	2.44	0.78	626
	35.6%	*1.95*	*0.83*	*87*
	2 = 49.5%			
	34.5%			
	3 = 35.3%			
	28.7%			
	4 = 6.4%			
	1.1%			
	5 = 1.0%			
	0%			
	6 = 0.2%			
	0%			
Integrating Sciences	1 = 18.8%	2.40	1.03	272
	*NA**	*NA**	*NA**	*NA**
	2 = 40.8%			
	3 = 25.4%			
	4 = 11.8%			
	5 = 3.3%			
	6 = 0%			

(table 3 cont.)

TABLE 3
Distribution of Science Scores
(Note: 1995-96 results are in plain text; *1994-95 results are in italics*)

Work Sample Description	Distribution by Score Level (%)	Mean (out of 6)	S.D.	Number
Laboratory Experiment	1 = 25.7% *30.5%*	2.03 *1.89*	0.80 *0.76*	1086 *502*
	2 = 50.0% *51.4%*			
	3 = 19.9% *15.7%*			
	4 = 4.1% *2.2%*			
	5 = 0.3% *0.2%*			
	6 = 0% *0%*			
Laboratory Observation	1 = 32.7% *NA**	1.87 *NA**	0.73 *NA**	1005 *NA**
	2 = 49.3%			
	3 = 16.7%			
	4 = 1.3%			
	5 = 0%			
	6 = 0%			
Literature Review and Evaluation	1 = 9.2% *13.2%*	2.63 *2.64*	0.87 *1.01*	790 *311*
	2 = 34.4% *31.8%*			
	3 = 42.5% *35.7%*			
	4 = 12.2% *16.1%*			
	5 = 1.6% *3.2%*			
	6 = 0% *0%*			

BEST COPY AVAILABLE

(table 3 cont.)

TABLE 3
Distribution of Science Scores
(Note: 1995-96 results are in plain text; *1994-95 results are in italics*)

Work Sample Description	**Distribution by Score Level (%)**	**Mean (out of 6)**	**S.D.**	**Number**
Societal Context of Science	1 = 16.5% *51.8%*	2.30 *1.65*	0.83 *0.82*	939 *112*
	2 = 44.2% *32.1%*			
	3 = 32.4% *13.4%*			
	4 = 6.5% *2.7%*			
	5 = 0.4% *0%*			
	6 = 0% *0%*			
Overall Portfolio Score	**Distribution by Score Level (%)**	**Mean**	**S.D.**	**Number**
Overall Science Portfolio	1 = 54.5%	1.52 *1.65*	0.61 *0.58*	1826 *553*
	2 = 39.3%			
	3 = 6.1%			
	4 = 0.1%			

* *In 1994-95, there was no Work Sample Description by this name.*

** *In 1994-95, there were so few entries corresponding to this Work Sample Description that meaningful indices could not be obtained.*

TABLE 4

Interrater Agreement and Correlations for Language Arts

Work Sample Description	N (25% of total)	Percent of Agreement Perfect/Adjacent/Resolved (1995-96 results are in plain text; *1994-95 results are in italics*)			Interrater Correl-ation
Analysis/Evaluation	358 *158*	65.6 *43.0*	33.2 *46.8*	1.1 *10.2*	0.81 *0.60*
Business and Technical Writing	59 _NA*_	69.5 _NA*_	30.5 _NA*_	0.0 _NA*_	0.86 _NA*_
Evaluation of Print or Electronic Media	46 *26*	73.9 *38.5*	23.9 *46.2*	2.2 *15.4*	0.69 *0.50*
Explanatory Writing	206 _NA*_	65.0 _NA*_	32.5 _NA*_	2.4 _NA*_	0.77 _NA*_
Imaginative Writing	520 *107*	61.7 *40.2*	34.2 *43.9*	4.0 *15.9*	0.78 *0.47*
Persuasive Writing	297 *70*	60.6 *34.3*	36.0 *55.7*	3.4 *10.0*	0.71 *0.47*
Proposing a Solution	59 *40*	49.2 *55.0*	50.8 *35.0*	0.0 *10.0*	0.75 *0.75*
Relating a Personal Experience	526 *89*	60.3 *47.8*	37.8 *45.6*	1.9 *6.7*	0.73 *0.79*
Research/Investigative Writing	288 *60*	64.2 *46.3*	33.7 *34.4*	2.1 *19.4*	0.84 *0.66*
Responding to a Literary Text	610 *73*	61.5 *39.7*	36.2 *46.6*	2.3 *13.7*	0.80 *0.55*
Writing about an Out-of-Class Reading	151 _NA*_	66.9 _NA*_	31.8 _NA*_	1.3 _NA*_	0.86 _NA*_
Writing about Uses of Language	25 _NA*_	64.0 _NA*_	32.0 _NA*_	4.0 _NA*_	0.68 _NA*_
Writing about Values/Issues/Beliefs	219 *47*	63.9 *60.7*	35.6 *28.6*	0.5 *10.7*	0.78 *0.72*
Writing a Review of the Visual or Performing Arts	85 _NA*_	58.8 _NA*_	40.0 _NA*_	1.2 _NA*_	0.80 _NA*_
Overall	734 *160*	74.3	25.6	0.1	0.77 *0.74*

* In 1994-95, there were so few entries corresponding to this Work Sample Description that meaningful indices could not be obtained.

BEST COPY AVAILABLE

TABLE 4 (cont.)

For Individual WSDs in **1995-96**:

Median Interrater Correlation: 0.78
Minimum Interrater Correlation: 0.68 (*Writing about Uses of Language*)
Maximum Interrater Correlation: 0.86 (*Writing about Out-of-Class Reading*)

For Individual WSDs in **1994-95**:

Median Interrater Correlation: 0.60
Minimum Interrater Correlation: 0.47 (*Imaginative Writing, Persuasive Writing*)
Maximum Interrater Correlation: 0.79 (*Analysis/Evaluation*)

TABLE 5

Interrater Agreement and Correlations for Mathematics

Work Sample Description	N (25% of total)	Percent of Agreement Perfect/Adjacent/Resolved (1995-96 results are in plain text; *1994-95 results are in italics*)			Interrater Correlation
Another Class	71	69.0	25.4	5.6	0.76
	*NA**	*NA**	*NA**	*NA**	*NA**
Analyzing Data	168	58.3	39.3	2.4	0.77
	*NA**	*NA**	*NA**	*NA**	*NA**
Connections	150	84.0	15.3	0.7	0.89
	*NA**	*NA**	*NA**	*NA**	*NA**
Collecting and Analyzing Data	140	68.6	28.6	2.8	0.70
	32	*90.6*	*9.4*	*0.0*	*0.96*
Consumer Beware	67	76.1	17.9	6.0	0.79
	*NA**	*NA**	*NA**	*NA**	*NA**
Comparing Notions	57	59.6	35.1	5.3	0.79
	80	*42.5*	*37.5*	*20.1*	*0.46*
Challenging Problem	266	63.5	30.8	5.7	0.79
	193	*77.7*	*10.9*	*11.4*	*0.58*
From Your Own Experience	60	63.3	35.0	1.7	0.82
	174	*79.9*	*10.9*	*9.1*	*0.60*
Logical Argument	100	51.0	31.0	18.0	0.58
	117	*47.9*	*35.0*	*17.1*	*0.61*
Multiple Methods	203	68.5	29.1	2.5	0.85
	160	*69.4*	*22.5*	*8.1*	*0.74*
Technology	126	61.9	34.9	3.2	0.86
	121	*75.2*	*17.4*	*7.5*	*0.73*
Overall	388	85.80	14.2	0.0	0.84
	120				*0.84*

** In 1994-95, there were so few entries corresponding to this Work Sample Description that meaningful indices could not be obtained.*

BEST COPY AVAILABLE

TABLE 5 (cont.)

For Individual WSDs in **1995-96**:

Median Interrater Correlation: 0.79
Minimum Interrater Correlation: 0.58 (*Logical Argument*)
Maximum Interrater Correlation: 0.89 (*Connections*)

For Individual WSDs in **1994-95**:

Median Interrater Correlation: 0.61
Minimum Interrater Correlation: 0.46 (*Logical Argument*)
Maximum Interrater Correlation: 0.96 (*Challenging Problem*)

TABLE 6

Interrater Agreement and Correlations for Science

Work Sample Description	N (25% of total)	Percent of Agreement Perfect/Adjacent/Resolved (1995-96 results are in plain text; *1994-95 results are in italics*)			Interrater Correlation
Applications	184 *46*	80.4 *58.7*	19.6 *32.6*	0.0 *8.7*	0.86 *0.51*
Design a Study	121 *43*	77.7 *53.5*	21.5 *44.2*	0.8 *2.3*	0.72 *0.54*
Design and Perform a Study	75 *33*	82.7 *36.4*	17.3 *54.5*	0.0 *9.1*	0.82 *0.35*
Evaluating Scientific Claims	211 *59*	73.9 *64.4*	25.6 *32.2*	0.5 *3.4*	0.75 *-0.04*
Historical Perspective	144 *25*	77.1 *40.0*	22.9 *48.0*	0.0 *12.0*	0.78 *0.00*
Integrating Sciences	66 *NA**	68.2 *NA**	31.8 *NA**	0.0 *NA**	0.77 *NA**
Laboratory Experiment	300 *156*	77.7 *55.8*	21.3 *41.0*	1.0 *3.2*	0.76 *0.55*
Laboratory Observation	204 *NA**	81.4 *NA**	18.1 *NA**	0.5 *NA**	0.76 *NA**
Literature Review and Evaluation	226 *93*	69.0 *33.0*	31.0 *53.8*	0.0 *12.9*	0.81 *0.44*
Societal Context of Science	317 *NA***	71.3 *NA***	28.4 *NA***	0.3 *NA***	0.74 *NA***
Overall	481 *139*	84.8	15.2	0.0	0.80 *0.38*

** In 1994-95, there was no Work Sample Description by this name.*

*** In 1994-95, there were so few entries corresponding to this Work Sample Description that meaningful indices could not be obtained.*

BEST COPY AVAILABLE

TABLE 6 (cont.)

For Individual WSDs in **1995-96**:

Median Interrater Correlation: 0.77
Minimum Interrater Correlation: 0.72 (*Design a Study*)
Maximum Interrater Correlation: 0.86 (*Applications*)

For Individual WSDs in **1994-95**:

Median Interrater Correlation: 0.44
Minimum Interrater Correlation: -0.04 (*Evaluating Scientific Claims*)
Maximum Interrater Correlation: 0.55 (*Laboratory Experiment*)

TABLE 7

Generalizability Analyses
(Note: 1995-96 results are in plain text; *1994-95 results are in italics*)

	Language Arts		**Mathematics**		**Science**	
Source	Variance Components	Percent of Total Variance	Variance Components	Percent of Total Variance	Variance Components	Percent of Total Variance
Student	0.3566	34.24%	0.5119	39.03%	0.1692	24.46%
WSD	0.0098	0.94%	0.1848	14.09%	0.1099	15.89%
Reader	0.0109	1.05%	0.0031	0.24%	0.0911	13.17%
Student* WSD	0.4419	42.42%	0.3504	26.72%	0.2805	40.55%
Student* Reader	0.0274	2.63%	0.1185	9.04%	0.0640	9.26%
WSD* Reader	0.0528	5.07%	0.1109	8.46%	-0.0475	0%
Error	0.1422	13.65%	0.0319	2.43%	0.0245	3.54%
G-coefficient	0.75 (1995-96) *0.73 (1994-95)*		0.79 (1995-96) *0.33 (1994-95)*		0.65 (1995-96) *0.31 (1994-95)*	